FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Wi-Fi Long Distance Maritime Communications Data Analytics

José Eduardo da Silva Timóteo de Carvalho

Master in Electrical and Computers Engineering

Supervisor: José Manuel Soares Oliveira

Second Supervisor: Manuel Alberto Pereira Ricardo

February 8, 2018

Resumo

Atualmente, as comunicações sem fios são cada vez mais importantes para o desenvolvimento da sociedade, não só em ambiente terrestre, mas também noutros contextos, como por exemplo o ambiente marítimo. Quando nos referimos às comunicações em ambiente marítimo, o cenário é completamente diferente do terrestre, sendo um ambiente mais difícil devido a vários fatores, tais como: o movimento na superfície do mar, as características da propagação de ondas rádio e a possível obstrução intermitente que diminui a eficiência da propagação do sinal.

Presentemente, existem várias soluções de comunicações sem fios em ambiente marítimo, tais como High Frequency and Very High Frequency, que no entanto não suporta altos débitos; as comunicações por satélite, que é uma tecnologia cara e não acessível à maioria dos utilizadores; e comunicações móveis (GSM, 3G e LTE), que apenas podem ser utilizadas perto da costa.

O objetivo principal desta dissertação é contribuir para a caracterização do canal de propagação e dos problemas associados à utilização da tecnologia Wi-Fi para diferentes frequências neste tipo de ambiente.

Nesta dissertação foi utilizada uma abordagem baseada em regressão para estabelecer relações entre variáveis externas e variáveis intrínsecas à transmissão de dados, contribuindo para um melhor conhecimento e caraterização do canal de transmissão em ambiente marítimo. Paralelamente, foi desenvolvida uma framework em R, intuitiva e automatizada, com o propósito de poder vir a ser utilizada em projetos futuros semelhantes.

Abstract

Nowadays, wireless communications are more and more important to the development of the society, not only in land, but also in the sea. When discussing about communications in maritime environments the scenario is different and harder, because of several factors, such as, the movement on the surface of the sea, the characteristics of the radio propagation and the possible intermittent obstruction that decrease the efficiency of signal propagation.

Plenty of wireless communications solutions are already used in maritime environment, such as HF/VHF, which doesn't support high rates; satellite communications, which is an expensive technology and not affordable for most of users; and mobile communications (GSM, 3G and LTE), that only ensure connection near the coast.

The main purpose of this dissertation is to contribute to the characterization of the propagation channel and the problems associated with the use of Wi-Fi technology for different frequencies in this kind of environment.

In this dissertation, a regression-based approach was used to establish relationships between external variables and variables intrinsic to data transmission, contributing to a better knowledge and characterization of the transmission channel in the maritime environment. At the same time, an automated and intuitive framework in R was developed with the purpose of being useful in similar future projects.

Acknowledgements

First, I would like to express special thanks to my supervisors, Professor José Manuel Soares Oliveira and Professor Manuel Alberto Pereira Ricardo for the opportunity they had conceded me to do this dissertation work, for the availability and support.

I would also like to thank the researchers of INESC-TEC involved with the MareCom project, Rui Campos and Carlos Leocádio, for their availability helping me with their expertise whenever I ran into a trouble spot or had a question about the project.

Thanks are also due to friends and university mates for moments of learning, support and solidarity. No personal achievement would be complete without them.

Last, but not least, I thank my family and girlfriend for all the support and words of encouragement they have provided during this journey and keeping me strong and advised during it.

José Eduardo da Silva Timóteo de Carvalho

"Start where you are. Use what you have. Do what you can."

Arthur Ashe

Contents

List of Figures

List of Tables

Abbreviations and Symbols

3G	Third Generation
AIC	Akaike's Information Criterion
AICc	Corrected Akaike's Information Criterion
BIC	Bayesian Information Criterion
BS	Base Station
CCS	Coordinated Centralized Scheduling
CDS	Coordinated Distributed Scheduling
CSMA/CA	Carrier-Sense Multiple Access with Collision Avoidance
CSV	Comma-Separated values
CTS	Clear to Send
CV	Cross-Validation
DATSA	Distributed Adaptive Time Slot Allocation
DCF	Distributed Coordination Function
DD	Decimal Degrees
DDM	Degrees Decimal Minutes
DIFS	Distributed Inter-Frame Space
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HF	High Frequency
HORST	Highly Optimized Radio Scanning Tool
IEEE	Institute of Electrical and Electronics Engineers
KML	Keyhole Markup Language
LTE	Long-Term Evolution
LOS	Line of Sight
MAC	Media Access Control
MAN	Metropolitan Area Network
MIMO	Multiple-Input Multiple-Output
MRPT	MAC-based Routing Protocol
MSE	Mean Squared Error
PCF	Point Coordination Function
PHY	Physical Layer
PR	Pattern Recognition
QoS	Quality of Service
RSS	Residual Sum of Squares
RSSI	Relative Received Signal Strength Indication
RTS	Request to Send
STA	Station
TCP	Transmission Control Protocol

TRITON	TRI-media Telematic Oceanographic Network
TSS	Total Sum of Squares
UDP	User Datagram Protocol
VHF	Very High Frequency
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network

Chapter 1

Introduction

1.1 Context

Nowadays, there has been an increasing number of terrestrial wireless networks. However, when discussing about communication in maritime environments, the scenario is completely different. This is due mainly because of the characteristics of this environment. For example, the movement on the surface of the sea, the characteristics of the radio propagation and the possible intermittent obstruction of the first zone of Fresnel [1].

There are some broadband communications used in the maritime environment, such as High Frequency and Very High Frequency (HF/VHF), which have the capacity to transmit over long distances but with low rates; satellite communications, which has a very high cost and most of the vessels do not have equipment installed for such use; and mobile communications, such as Global Positioning System, 3G and Long-Term Evolution that only ensure connection near the coast.

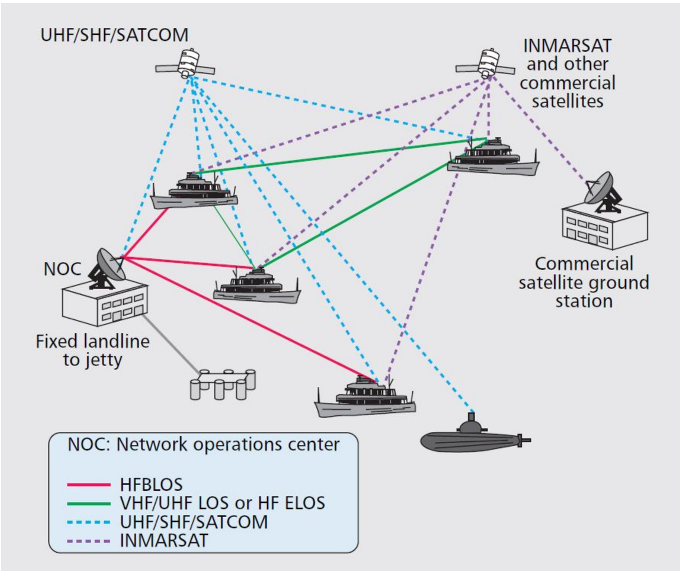


Figure 1.1: Typical maritime network [1].

[1] A Fresnel zone is a cylindrical ellipse drawn between transmitter and receiver.

1.2 Motivation

Most of the communities operating in a maritime environment, such as fishermen, shipping companies, recreational craft and military vessels (Navy), are increasingly requiring a highly available, broad band and low cost maritime communications solution.

The availability of wireless, mobile and long-range communications, that are capable of serving fixed and mobile floating platforms, located in remote areas of the ocean, which are typically beyond the reach of terrestrial networks, would undoubtedly add value to the maritime communities.

1.3 Objectives

The purpose of this thesis is to contribute to one of the main objectives of the MareCom project, which is the complete characterization of the propagation channel and the performance/problems associated with Wi-Fi technology and its variants (IEEE 802.11a/g/n/ac/ah) for different frequency bands, such as 700 MHz, 868 MHz, 2.4 GHz and 5.8 GHz, in a maritime environment.

The MareCom project [2] is led by Wavecom, with INESC TEC being a partner together with the Portuguese Navy and Ubiwhere. This project envisages the development of a solution that is an alternative to the satellite and VHF radio technologies that are currently used, serving this way the communities that operate in maritime environment, for example, fishing fleets or maritime transport, with reliable and broadband links and integration with the ground communications scenario, increasing the digital inclusion of those communities.

Specifically, this work intends to search for and identify relationships in large amounts of information (Big Data) related to the transmission of data in maritime environments, especially the relationship between data communication variables and external variables, such as, environmental variables, time, location, direction or operation of the vessel, enabling the identification of clusters of data corresponding to typical transmission contexts.

Another of the main objectives of this thesis is to develop an automated and intuitive framework that can be used in similar future projects that might arise. The major purpose of this framework is to calculate some parameters that help to understand some aspects, such as the distance traveled by the receiver or what influence some communication or external variables may have in the quality of the connection, using the data generated during the tests.

1.4 Document structure

This document is divided into seven chapters, corresponding the first to the present chapter that serves as an introduction to the developed work and its motivation. In Chapter 2, we describe the state of the art and identify some related works in the area of maritime communications. We also describe some data analytics techniques that can be used to analyze maritime data communications. Chapter 3 describes in more detail the data analytics method adopted during the dissertation. In Chapter 4, a description of the network topology, the hardware and the software used in the MareCom project testbed is made. Chapter 5 describes the data processing approach in order to search for and identify relationships in large amounts of information related to the transmission of data in maritime environments. The data analytics using regression is presented in Chapter 6. An analysis of the data is made taking into account various aspects of maritime communication in order to understand the impact of this type of environment on Wi-Fi communications. Lastly, in Chapter 7, the final results are briefly discussed, ending with future work directions. An appendix is also added, presenting the relevant R and python code used in the scope of this work.

Chapter 2

State Of The Art

The purpose of this chapter is to present recent existing solutions that implement maritime broadband communications, including protocols and maritime propagation models. The chapter also describe some past maritime communications experiments and scenarios. In addiction, it also refers some data analytics techniques that can be used to analyze experimental data obtained from maritime communications testbeds.

2.1 Maritime communications

In this section, firstly we detail the maritime communications environment and afterwards we present past maritime communications experiments, developed protocols and propagation models suitable for this environment.

2.1.1 Maritime communications environment

Unlike terrestrial networks, in maritime wireless networks the received signal strength are subjected to perturbations due to the sea movement. The main point is not how to build new technologies, but how to adapt existing ones to handle with specific maritime communications characteristics.

Since it is almost impossible to predict maritime conditions, people usually face weak and unstable communications when they are in the sea. The movement of the sea caused by sea waves affects ships with communications equipments, causing variations to the communication antenna's height and orientation, affecting the antenna gain and the overall quality of the link. This makes the quality of maritime connection link experience periodic degradation, increasing the number of packets retransmissions and provoking long delays.

Figure 2.1 illustrates two different ship movements that cause variations in the received signal strength. Note that the variations of the antenna's tilt change significantly more the antenna's gain comparing to the variations in antenna's height.

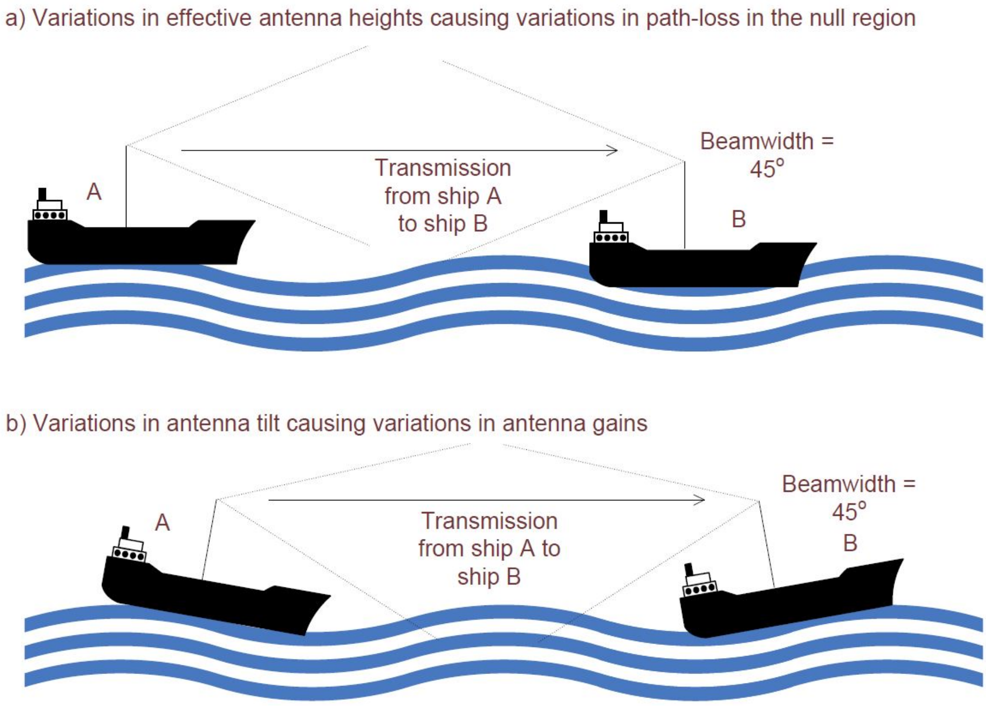


Figure 2.1: Variations in received signal strength provoked by sea waves [3].

2.1.2 802.11

IEEE 802.11 standard targets Media Access Control (MAC) and Physical Layer (PHY) specifications in order to implement Wireless Local Area Network (WLAN) computer communications with different frequency bands [4]. The MAC layer manages and maintains communications and it is based on listen-before-send techniques, such as Career Sense Multiple Access with Collision Avoidance (CSMA/CA), which detects collisions, avoids them by transmitting only when the channel is free. This standard specifies two forms of medium access, Distributed Coordination Function (DCF) and Point Coordination Function (PCF). DCF is mandatory to be implemented in all stations. With DCF, 802.11 station shall sense the medium and check if any other station is sending a frame before transmitting. The medium can be idle or busy. If the medium is busy, the station must delay the transmission to avoid collisions. If the medium is idle, for a given interval (Distributed Inter-Frame Space) the station is able to transmit. In case of a busy medium, after the delayed time by the station, before attempting to transmit, the station must select a random back off time interval till the medium is idle again. After the medium is idle, and both time are awaited (delayed and back off time), the sender transmits a Request to Send (RTS) and waits for a Clear to Send (CTS), as exposed in Figure 2.2. If the CTS is acquired, the sender is able to transmit.

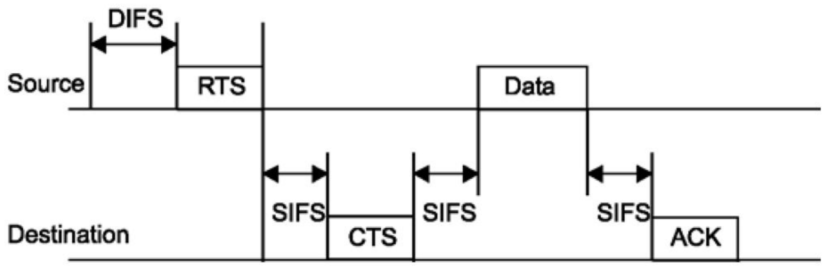


Figure 2.2: RTS/CTS/Data/ACK setting [5].

IEEE 802.11n, which is an enhanced standard of IEEE 802.11a and IEEE 802.11g [6], has emerged because of the requirement of transmitting higher bit rate. Due to the use of Multiple-Input Multiple-Output (MIMO) technique, IEEE 802.11n increases maximum net data rate from 54 Mbit/s to 600 Mbit/s. The use of MIMO gives many benefits, such as antenna diversity and spatial multiplexing and has the capability to synchronously resolve information from multiple signal paths using spatially separated receive antennas [7] as shown in Figure 2.3.

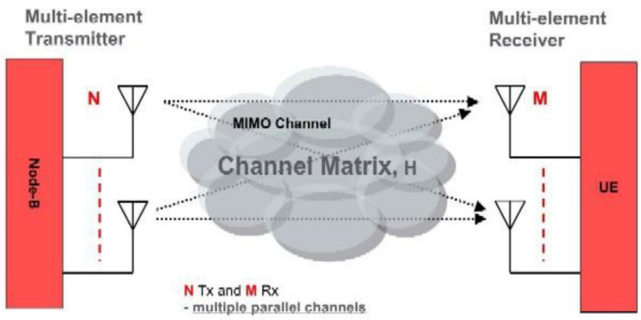


Figure 2.3: Multiple In Multiple Out [8].

According to IEEE 802.11n, the antennas configuration is defined by AxB:C, where A is the maximum number of transmit antennas, B is the maximum number of receive antennas and C is the maximum number of data spatial streams [9]. Common configurations are 2x2:2, 2x3:2 and 3x2:2, being able to go up to 4x4:4, which enables net data rate till 600 Mbit/s. The spatial multiplexing technique fragments the information to be sent into spatial streams, taking in consideration the several antennas available to transmit. Since every spatial stream has its own distinctive signature, upon reaching the receiver they are assembled neatly. The maximum number of spatial streams is limited by the minimum number of antennas on transmitter or receiver. In order to reach the maximum throughput of 600 Mbit/s, there are some aspects to take in consideration. It is required to exist 4 transmit antennas and 4 others receive antennas allowing to be used 4 spatial streams. It is also necessary that the transmitter and the receiver be near (tens of meters), so that the guard interval of 400 ns allows to transmit information without loss of packets. The guard interval exists to ensure that there is no interference in receiving data, which could happen when receiving two different data blocks transmitted in different times intervals, one received through the direct ray and the other of a signal reflection [10].

As mentioned in Section 2.1.1, it is quite difficult to provide communications for ships beyond the cellular coverage. A technology that can be used to overcome this problem is mesh network technology. In IEEE, there are quite a few standards that address mesh networking technology. Unfortunately, in maritime environments the appliance of these standards isn't straightforward.

IEEE 802.11s is a mesh network modification to the IEEE 802.11, which uses the CSMA/CA for channel access [11]. In agreement with the analysis based on ship traffic movement data obtained from the Maritime Port Authority of Singapore (MPA) [12], in a maritime environment as shown in Figure 2.4, with the purpose of forming a well-connected mesh network, the transmission should be capable of reaching at least 18 km. Since IEEE 802.11s supports only ranges of up to several hundred meters, it is not applicable for maritime mesh communication. The following section presents WiMAX, a standard that overcomes this range problem, being more suitable to maritime communications.

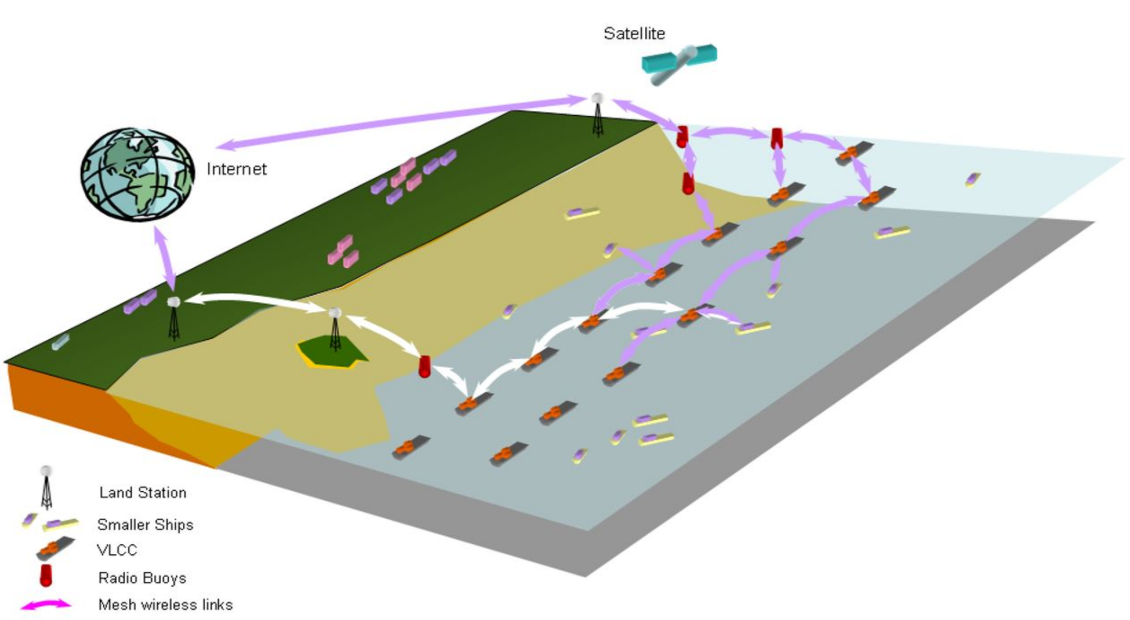


Figure 2.4: Maritime mesh network [12].

2.1.3 WiMAX

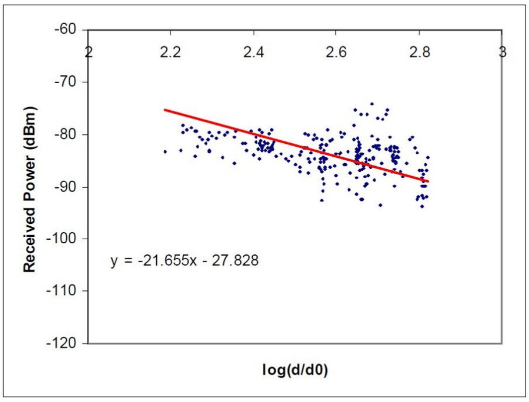
IEEE 802.16 is also often referred to as WiMAX and defines a mesh network standard for a wireless Metropolitan Access Network (MAN) [13].

This protocol supports interoperable air interfaces from 2 to 66 GHz with a common MAC layer and is able to cope with user mobility up to a vehicular speeds. This standard is able to deal with functions required for a mobile broadband wireless access system, such as network discovery/selection, network entry and exit, Quality of Service signaling and management, security, mobility (handover management), power management modes between a client device and a Base

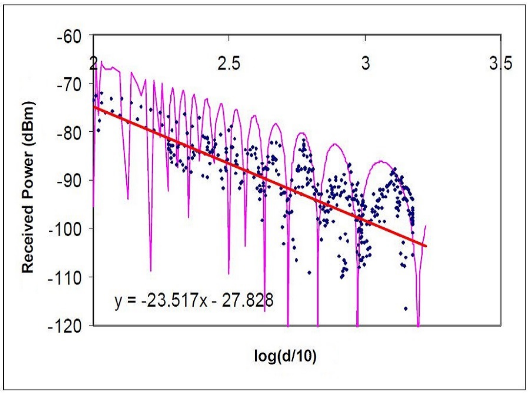
Station (BS). As stated in Section 2.1.2, IEEE 802.11s is also a mesh network but it is not suitable for the maritime communication, since it doesn't support a range of several tens of kilometers, while IEEE 802.16 supports a large radius coverage, which may vary from few kilometers to about 50 km and high data rate with cheaper access.

There are several performance experimental studies regarding WiMAX protocol in maritime scenario. In order to detect how the pathloss changes according to antennas height, few experimental measurements were performed using WiMAX at 5.8 GHz in Singapore Port [14]. Three distinct scenarios were performed for several transmitter antenna heights (4m, 76m and 185m). The output signal was 30 dB using an antenna, which had omni directional and with 12 dBi gain. In the receiver, the same antenna was used and was placed on a diving boat, 8 m from sea surface.

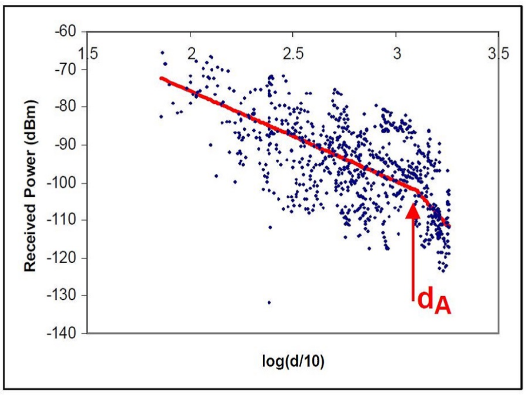
In Figure 2.4(a), it is possible to understand that, considering a normalized distance between 2.2 and 2.5, the received power diverse less, compared to a normalized distance bigger than 2.5. Since distances bigger than 2.5 correspond to the ship's anchorage area, it is possible to detect a bigger variation due to reflections from large ships in this region, while in the region between 2.2 and 2.5 the receiver had a good Line of Sight (LOS) to the transmitter.



(a) Transmitter antenna 185m height.



(b) Transmitter antenna 76m height.



(c) Transmitter antenna 4m height

Figure 2.5: Received power for different antenna heights [14].

In Figure 2.4(b), the transmitter antenna was mounted in a light house and it is much likeness to the two-ray path model, where the first ray was a LOS signal from the transmitter to the receiver while the second one was a reflected signal from the transmitter to the sea surface and afterward to the receiver. With frequency at 5.8 GHz, the sea surface satisfies good conductor condition.

In Figure 2.4(c), the transmitter antenna was placed on a tripod about 4m from the sea level. Contrary to the other previous scenarios, the LOS wasn't dominant during the measurement. It is possible to check a breakpoint dA. When the distance is bigger than dA the signal decreases really fast, limiting the coverage zone of WiMAX. It is also possible to detect a high divergence of the received power due to the transmitter location.

2.1.4 TRITON

The TRI-media Telematic Oceanographic Network (TRITON) project aims to develop a system for high-speed and low cost maritime communications and was implemented in East Coast of Singapore port [15]. This project investigates a wireless mesh network based on IEEE 802.16 standard at 5.8 GHz and 2.4 GHz. The IEEE 802.16 is capable of supporting a much longer transmission range, good QoS and a longer communications distance, which is an advantage for maritime communications. This solution is essentially made to regions with high ship density. Facing the unique characteristics of maritime mesh networks, in TRITON some protocols on MAC and network layer have been developed and enhanced.

The enhancements in the MAC layer are:

- Operates with Coordinated Distributed Scheduling (CDS), rather than Coordinated Centralized Scheduling (CCS) scheme, which is more advisable for a dynamic maritime mesh network.

- Developed a new algorithm for resource allocation: Distributed Adaptive Time Slot Allocation (DATSA), allowing a mesh node to keep a track of not only the availability of its own but also of the neighbor's data resource. Since the nodes are constantly moving, it causes a resource overlap, which this algorithm can also detect.

- Developed a multi-channel transmission algorithm, since in a maritime mesh network a single frequency channel isn't enough to provide sufficient bandwidth.

- Designed protocols to resolve fairness problems, which exists in 802.16 mesh network.

In Network layer, a new routing protocol was developed, the MAC-based Routing Protocol for TRITON (MRPT), which is an optimization of the typical routing protocols. MRPT is a proactive routing protocol. It manages multiple routing paths and uses mesh MAC control messages to transmit routing information from the land station to the ships, reducing the initial packet delay.

The main point of TRITON architecture is to achieve the maximum range by designing a multi hop mesh network with buoys, neighboring ships and maritime beacons. The ships are connected to terrestrial networks by land stations placed along the coast, while buoys are positioned to extend the connectivity and relay traffic when necessary. Figure 2.6 shows the high level architecture of TRITON.

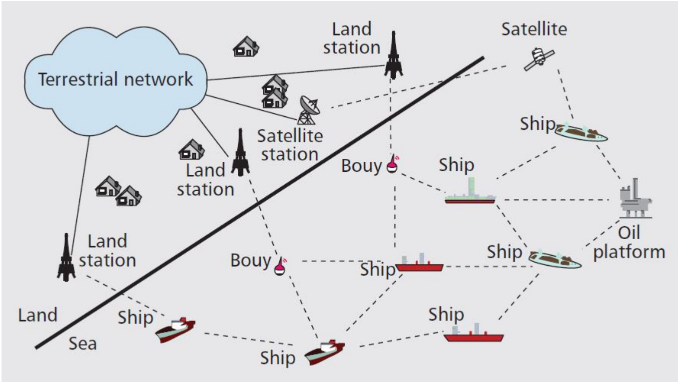


Figure 2.6: High level architecture of TRITON [16].

2.1.5 Propagation models

It is important to keep in mind that the land radio propagation is different compared with channel propagation in the sea due to refractions, reflections and constant oscillations. In [17], a few measurements using a 2.5 GHz mobile WiMAX test-bed in sea port scenario were made to evaluate their influence on communications. This article stated that a 2-ray pathloss model [18] (direct ray and reflected one) is able to fit the behavior of the observed maritime wireless channel. The simulation were made using the OPNET[1] simulator and, although not being able to predict the behavior in all situations, it could reproduce in most cases the system performance.

Figure 2.7 illustrates the radio-propagation environment. In order to proceed the measurements, a BS was placed 30 meters above the sea surface with a gain of 17 dBi transmitting at 35 dBm of power and the boat was equipped with a 11 dBi external omni-directional antenna set at 10 meters of the sea surface, attached to a laptop in order to get time synchronization and geographical position.

The system performance was evaluated in terms of RSSI and throughput at IP level. To avoid the congestion algorithm's limitation of Transmission Control Protocol (TCP), User Datagram Protocol (UDP) was used.

Several measurements of RSSI were made for different distances, which is represented in Figure 2.8 .

In Figure 2.8, due to the characteristics of maritime environment, it is possible to observe some deep holes up to 5 km distance, which stands for a breakpoint. After this breakpoint, RSSI seems to be quite stable with a linear decrease of less than 1 dB/Km and it continuously degrades. After approximately 19 km far from the shore, the connection broke down.

[1]OPNET Technologies, Inc. is a business software that provides performance analysis for computer networks and applications.

Figure 2.7: Site of installation [17].

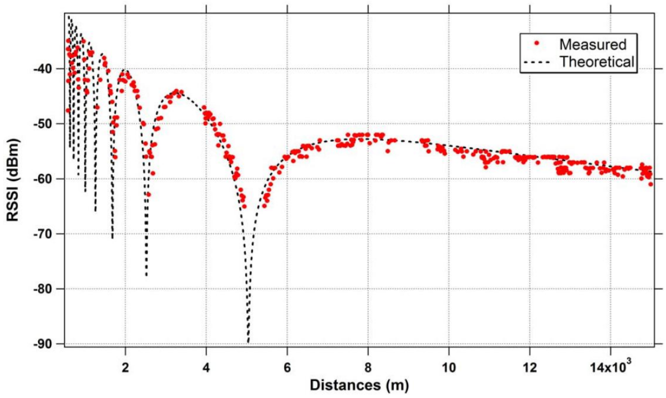


Figure 2.8: Theoretical and measured RSSI for different distances [17].

Based on measurements, a 2-ray propagation model was developed:

Power Received:

$$P_r = \frac{P_t G_t G_r}{L_2 ray} \tag{2.1}$$

Proposed 2-ray Pathloss Model:

$$L_2 ray = \frac{L_f s}{\beta} \tag{2.2}$$

Reflection Coefficient:

$$\Gamma(\theta_i, n_1, n_2) = \frac{n_1 cos\theta_t - n_2 cos\theta_i}{n_1 cos\theta_t + n_2 cos\theta_i} \tag{2.3}$$

$$\theta_t = arcsen\left(\frac{n_1}{n_2} sen\theta_i\right) \tag{2.4}$$

$$\beta = 1 + \Gamma(\theta_i, n_1, n_2)^2 - 2\Gamma(\theta_i, n_1, n_2) cos\left(\frac{4\pi h_t h_r}{\lambda d}\right) \tag{2.5}$$

Free Space Pathloss Model:

$$L_f s = \left(\frac{4\pi d}{\lambda}\right)^2 \tag{2.6}$$

where P_t, G_t and G_r represents transmission power, transmitter antenna gain and receiver antenna gain, respectively. The h_t and h_r represent the height of the transmitter and receiver antenna, respectively. The $L_f s$ corresponds to the free space pathloss model, where λ is the wavelength and d is the distance between Mobile Station (MS) and a BS. θ_i represents wave angle of incidence, while θ_t is the angle of transmitted wave, $n_1 \cong 1$ and $n_2 \cong 1.333$ represents refraction index of air and water, correspondingly.

Analyzing both models, free space model and 2-ray model, we are able to identify considerable similarities, except the 2-ray pathloss peaks, as shown in Figure 2.9.

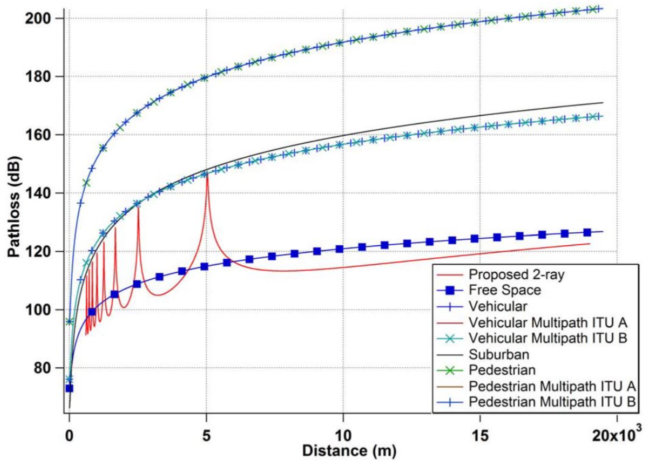


Figure 2.9: Simulated pathloss models results [17].

2.2 Data analytics

Data analytics is usually associated to the idea of building a statistical model for predicting/estimating an output based on one or more inputs with the goal of discovering useful information.

Data analytics is often designated as data mining. Data mining is a broad term that usually means the detection of patterns in large data sets, using methods from statistics, machine learning and database systems. It is divided in three main steps:

1. Selection and pre-processing of data.
2. Application of a methodology, such as regression, classification or clustering, that will be explained in the following sections.
3. Analysis and validation of results.

Typically, data analytics problems are divided in two categories: supervised or unsupervised. In the supervised category, there is a correspondence between each observation of the predictor measurement(s) $x_i, i = 1, \ldots, n$ and a response measurement y_i, while in the unsupervised category, for every experience $i = 1, \ldots, n$, we observe a vector of measurements, instead of an associated response y_i.

In relation to variables, they can be defined as quantitative or qualitative. Quantitative variables correspond to numerical values, for example, person's age, price of the stock, the value of an item, etc. On other hand, qualitative variables take on values in one of K different categories or classes, for example the brand of a product or the eye color. While problems with a quantitative response are often referred as regression problems, qualitative problems are often related to classification problems.

Actually, data analytics is really useful and interesting for several situations, such as, identify anomalous information, identify information patterns, analyze fraud, etc.

These techniques that are presented in the following sections, have been widely used also in networks and telecommunications fields, as for example in [19], [20] and [21].

2.2.1 Regression

Linear regression is a simple approach for supervised learning and is a statistical process for estimating the relationships among variables [22].

Usually, linear regression is used for the following activities:

- Prediction - To predict a future response variable based on known values of the predictor variables;
- Description - To measure the effect of changing a controllable variable on the mean value of the response variable;
- Control - To confirm that a process is providing expected results under the present operating conditions.

2.2.2 Classification

As mentioned in Section 2.2, qualitative variables are often related to classification methods, i.e., classification is an approach for predicting qualitative responses. Normally classifying is predicting a qualitative response for an observation, while appointing the observation to a category or class.

There are three most used classifiers: linear discriminant analysis, logistic regression and K-nearest neighbors [23].

A set of observations $(x_1, y_1), \ldots, (x_n, y_n)$ are made, in order to build a classifier. A good classifier is one that has a good performance on the training data and on test observations that were not used to train the classifier.

2.2.3 Clustering

Clustering refers to several set of techniques for finding clusters in a data set, in other words, it is the process of discovering groups (clusters) in the data that are in somehow similar. It is a common technique for data analytics, used in many fields, such as pattern recognition, image analysis, machine learning and many more, which leads to a great number of clustering methods. The two best-known clustering techniques are: K-means clustering and hierarchical clustering.

2.2.3.1 K-Means clustering

In this approach, the observations are partitioned into a pre-specified number of clusters. This technique requires two main steps:

1. Define the desired number of clusters K.

2. Specifically one of the K clusters will be assigned to each observation by the K-means algorithm. Figure 2.10 illustrates the results performing a K-means clustering with 150 observations in two dimensions, and three distinct values of K. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm.

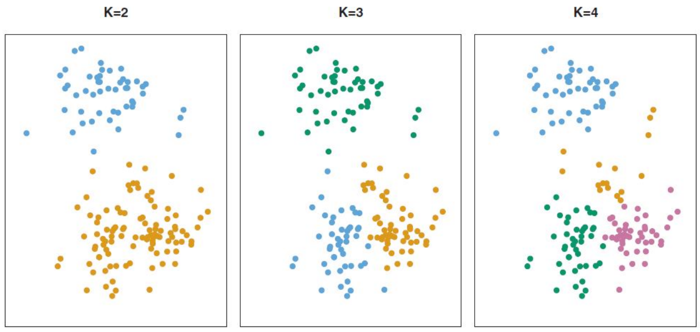


Figure 2.10: K-means clustering with 150 observations in two dimensions, and three distinct values of K. [23].

2.2.3.2 Hierarchical clustering

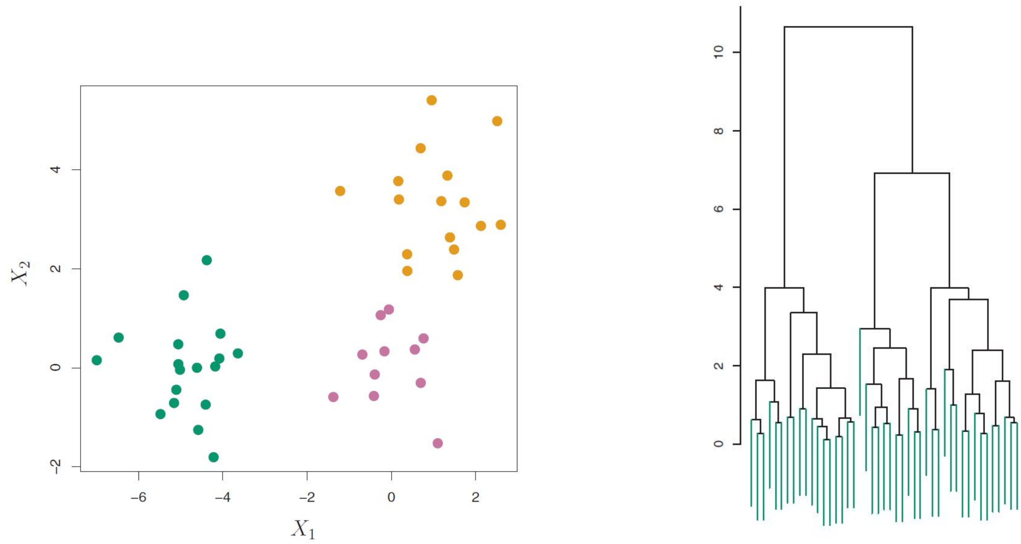
For the previous method, it is necessary to pre-specify the number of clusters K, which is an disadvantage, since usually the number of clusters we want is unknown. Contrary to K-means, hierarchical clustering does not require that we choose a particular K. In this approach, a tree-like visual representation of the observations allows us to check at once the clusterings obtained for each possible number of clusters, from 1 to n. In other words, it is a diagram used to illustrate the arrangement of the clusters, named dendrogram.

2.2.4 Pattern recognition

This method focus on the recognition of patterns in data through the use of computer algorithms, more precisely Pattern Recognition (PR) is the assignment of some output value to a given input value, according to a certain PR algorithm [24].

There are two types of pattern recognition methods: supervised learning and unsupervised learning. Supervised learning takes on that a set of training data has been provided, consisting of a set of instances that have been properly labeled with the appropriate output, while unsupervised learning on the other hand, assumes that the data has not been labeled and tries to find inherent patterns, in order to determine the correct output value for new data instances.

Basically, pattern recognition is a computational algorithm used to classify raw data and it has few methods, that are used on the development of several applications in different fields. Examples of pattern recognition methods are: statistical pattern recognition, syntactic pattern recognition, structural pattern recognition and neural pattern recognition.



(a) Forty-five observations generated in two-dimensional space.

(b) Dendrogram obtained from hierarchically clustering.

Figure 2.11: Hierarchical clustering and corresponding dendrogram [23].

2.2.5 Outlier detection

An outlier is a data point that is notably different from the remaining data. Hawkins defined [25] the view of an outlier as follows: "An outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."

Outliers are often designated as abnormalities or anomalies in the data mining and happens when the generating process operates in an unusual way [26].

The definition of outlier detection is: given a data matrix D, the rows of the data matrix that are very different from the remaining rows are determined. Nowadays, there are plenty of several applications that use this method, such as, intrusion-detection systems, credit card fraud, medical diagnosis, sensor events, etc.

2.2.6 Association rules

This is a method for discovering interesting relations between variables in large databases. In other words, association rules indicate a statical correlation between the occurrences of certain data items in a data set [27].

The generic form of an association rule is

$$X_1, \ldots, X_n => Y[C, S], \tag{2.7}$$

where the attributes $X_1, \ldots, X_n$ predict Y with a confidence C and significance S.

Since it was one of the first algorithms used in data analytics, it has serious limitations, like they normally assume the data consists of only numeric and textual symbols, not containing, for example images, and it usually assumes the data was fully collected from a single database, with the purpose of a data mining task in mind.

Chapter 3

Data Analytics Using Regression

Regression analysis is usually used for prediction and forecasting, being its use in machine learning quite often. Regression analysis is also used to understand which independent variables are related to the dependent variable, and to explore the forms of these relationships.

In this dissertation, the regression analysis was applied with the purpose of identifying relationships between data communication variables and external variables and identify relationships related to the transmission of data in maritime environments.

3.1 The linear model

The purpose of linear regression is to model a continuous variable Y as a mathematical function of one or more X variable(s), so that we can use this regression model to predict Y when we only know X.

In the context of maritime communications and particularly in our study, it is expected that several variables, such as distance, accelerometer, gyroscope, noise, packet length, etc, have some influence in the received power (RSSI). Applying this model, we intend to answer some important questions, such as:

1. *Which variables have influence in the RSSI?*
   Do all variables contribute to RSSI or just some of them? To answer this question, we must find a way to separate the individual effects of each variable.

2. *Is there a relationship between RSSI and distance and how strong is this relationship?*
   It is expected that distance has influence in RSSI. Assuming that there is a relationship between them, we would like to know the strength of this relationship. In other words, if the distance is the main variable affecting the RSSI.

3. *Does the ripple oscillation influence the RSSI?*
   Since we are in a maritime environment, does the sea ripple affects RSSI? What is the strength of this relationship?

3.1.1 Simple linear regression

In the most elementary case, there is a linear relationship between the forecast variable (Y) and a single predictor variable (X),

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{3.1}$$

The parameters β_0 and β_1 correspond to the intercept and the slope of the line, respectively. The intercept β_0 represents the predicted value of Y when X is 0, while β_1 represents the average predicted in Y resulting from a one unit increase in X. ε is the error term. Figure 3.1 represents an example of data from such a model.

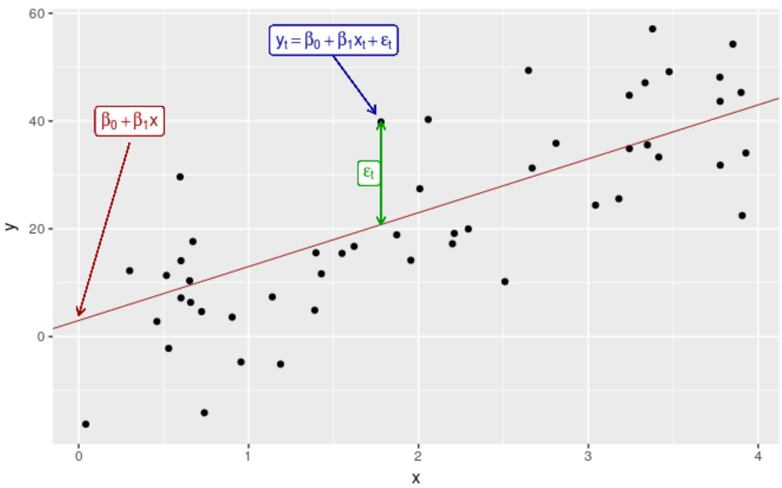


Figure 3.1: An example of data from a linear regression model [28].

We can consider that each observation Y consists of the systematic part of the model $\beta_0 + \beta_1 X$ and the random error, ε. The ε term does not mean an error, but a deviation from the underlying straight line model.

3.1.2 Multiple regression

Normally, in practice we have more than one predictor. For example, in our data set we have several predictors that might have some influence in RSSI, such as: the distance, the accelerometer, the noise, or the packet length.

One approach is to run simple linear regression, using a different predictor variable at a time. However this option is not absolutely satisfactory.

Instead of fitting a independent simple linear regression model for each predictor, a better approach is to extend the simple linear regression (3.1) so that it can directly contain multiple predictors. When there are two or more predictor variables, the model is called multiple regression and is given by the equation (3.2)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon, \tag{3.2}$$

where Y is the variable to be forecast and $X_1, \dots, X_k$ are k predictor variables. After taking account of the contribution of all predictors, the coefficients $\beta_1, \dots, \beta_k$ measure the effect of each predictor.

3.2 Estimating the coefficients

In practice, we have a collection of observations but we do not know the values of the coefficients β_0, β_1, in case of simple regression. These coefficients need to be estimated from the data. Our goal is to obtain coefficients estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, such that the linear model fits the available data. Specifically, we want to find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ such that the resulting line is as close as possible to the n data points. There are plenty of ways to estimate coefficients, however the most common approach involves minimizing the sum of the squared errors.

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction of Y based on the ith value of X. The ith residual is represented by $e_i = y_i - \hat{y}_i$. We define the residual sum of squares (RSS) as

$$RSS = e_1^2 + e_2^2 + \dots + e_n^2, \tag{3.3}$$

or identically as

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \tag{3.4}$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.

3.3 Measuring the quality of fit

A way to measure how well the predictions match the observed data is required, in order to evaluate the performance of a statistical learning method. In other words, we need to quantify the amount of data to which the predicted response value for a given observation is close to the true response value for that observation. The most commonly-used measure in the regression setting is the Mean Squared Error (MSE) given by

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2, \tag{3.5}$$

where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the ith observation. If the predicted responses are very close to the true responses, the MSE will be small. On other way, it will be large if, for some observations, the predicted and the true response differ considerably.

3.4 Selecting predictors

When there are several predictors, we need a strategy to select the best predictors to use in a regression model. The main question we want to address when selecting the predictors is:

- *Do all the predictors contribute to explain Y, or only a subset of them are useful?*

  It is possible that all predictors are associated with the response, but usually the response is only related to a subset of predictors. The process of determining which predictors are associated with the response is called variable selection.

  In order to perform an ideal variable selection, we try out a lot of different models, each containing a different subset of predictors. For example, if we have two predictor variables, then we can consider four models: (1) a model containing no variables, (2) a model containing X_1 only, (3) a model containing X_2 only and (4) a model containing both X_1 and X_2. After the models are produced, another question arises: how do we determine which model is best? Several statistics can be used to judge the quality of a model, some of them will be presented in the following sections.

There are a total of 2^p models that contain subsets of p variables, which means that, even for moderate p, trying out every possible subset of the predictors is impractical. Taking in account our data set, the number of predictors is small, making it possible to consider all the 2^p models.

There are several measures to select predictors, such as cross-validation, the Akaike's information criterion (AIC), the corrected Akaike's information criterion (AICc), the Bayesian information criterion (BIC) and the adjusted R^2 that will be introduced in the following subsections.

3.4.1 Cross-Validation

The Cross-Validation (CV) process uses the following steps:

1. Remove observation n from the data set, and fit the model using the remaining data.
2. Compute the error $e_n^* = y_n - \hat{y}_n$ for the omitted observation. This is not the residual error presented in section 3.2, because the nth observation was not used to estimate the value of $\hat{y}_n$.
3. Repeat step 1 and 2 for $n = 1, \ldots, N$.
4. Compute the MSE from $e_1^*, \ldots, e_N^*$.

Earlier, performing a CV was almost impossible for many problems with many predictors and/or large number of observations, causing AIC, BIC and adjusted R^2 a more inviting approaches. However, nowadays with fast processing computers, the CV is a very attractive approach.

Under this criterion, the best model is the one with the smallest value of CV.

3.4.2 Akaike's information criterion

AIC compares the quality of a set of statistical models to each other. Given a collection of models for the data, AIC estimates the quality of each model, providing in this way a mean for model selection.

Although the AIC will choose the best model from a set, it won't say anything about the absolute quality of the best model. In other words, AIC won't give any warning if all models are poor.

This method is defined as

$$AIC = N \log\left(\frac{RSS}{N}\right) + 2(k+2), \tag{3.6}$$

where k is the number of predictors in the model and N is the number of observations used for estimation. The $k+2$ term occurs because there are $k+2$ parameters in the model: the k coefficients for the predictors, the variance of the residuals and the intercept. Although, AIC has slightly different definitions in distinct computers packages, they should all lead to the same model being selected. This method idea is to penalize the fit of the model (RSS) with the number of parameters that need to be estimated.

Usually, the model with minimum value of AIC is often the best model.

3.4.3 Corrected Akaike's information criterion

The AIC tends to select too many predictors for small values of observations. Basically, AICc is AIC with a correction for finite sample sizes. The formula for AICc is as follows:

$$AICc = AIC + \frac{2(k+2)(k+3)}{N-k-3}. \tag{3.7}$$

Such as AIC, the AICc should be minimized.

3.4.4 Bayesian information criterion

The BIC, is related to the AIC. The method is defined as

$$BIC = N \log\left(\frac{RSS}{N}\right) + (k+2)\log(N). \tag{3.8}$$

The BIC criterion will choose the same model as AIC criterion. The reason why this happens, is that BIC penalizes the number of parameters more heavily than AIC.

As AIC, the model with the lowest BIC is preferred.

3.4.5 Adjusted R^2

The R^2 is a statistical measure of how close the data are to the fitted regression line.

The R^2 is defined as $1 - RSS/TSS$, where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares for the response. Since the larger the number of variables the smaller the RSS, the R^2 always increases as more variables are added, even if that variable is irrelevant.

R^2 is not a satisfactory measure of the predictive ability of a model, because if a model which generates forecasts that are precisely 20% of the actual values, the R^2 value would be 1 (perfect correlation), but the forecasts are not very close to the actual values.

To overcome these problems, the Adjusted R^2 has been created and nowadays it is a popular approach for selecting among a set of models that contain different numbers of variables. The Adjusted R^2 also called "R-bar squared" is defined as

$$R^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1}, \tag{3.9}$$

where N is the number of observations and k is the number of predictors. With this upgrade, the R^2 will no longer increase with each added predictor.

Unlike cross-validation, AIC, AICc and BIC, for which a small value indicates a better model, a large value of adjusted R^2 indicates a better model.

3.5 Regression trees

A regression tree is built through a process known as binary recursive partitioning. The models are achieved by recursively partitioning the data space and fitting a simple prediction model within each partition, represented graphically as a decision tree [29]. Decision trees analysis is usually done when the predicted outcome can be considered a real number.

In our study, we will use the regression trees to illustrate the relationship and the influence between the communication and external variables. With these regression trees, it is possible to quantify the values of outcomes and the probabilities of achieving them, enabling to draw conclusions about the study.

Chapter 4

Experimental Setup

As stated before, this dissertation aims to analyze the generated data in MareCom project. This chapter describes the network topology, the hardware and the software used in the project testbed.

4.1 Network topology

Figure 4.1 details the network topology used to perform the tests.

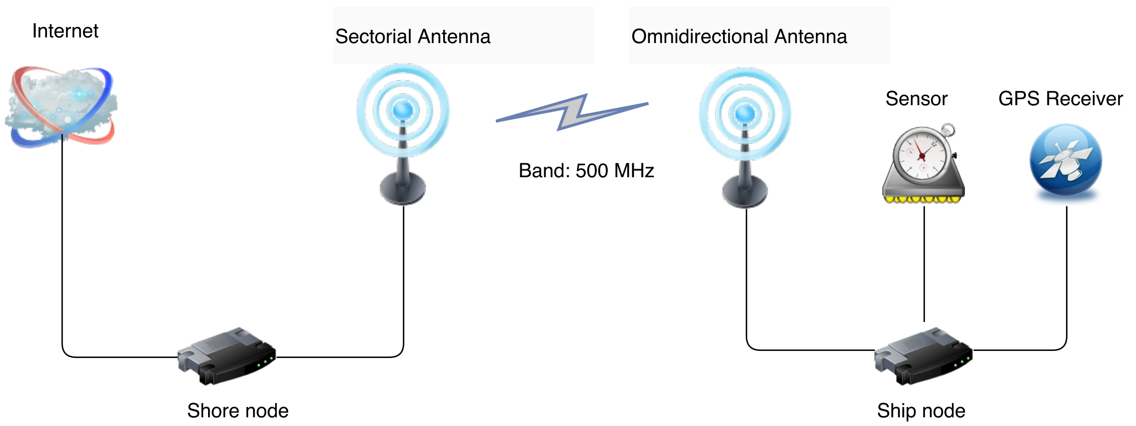


Figure 4.1: Network topology.

The shore node is the only one with Internet access via a 3G connection, in order to allow remote access. The 3G connection allows remote monitoring and configuration of equipment, as well as the download of logs resulting from the tests.

4.1.1 Hardware

The hardware used in the testbed is described in Table 4.1.

Table 4.1: MareCom testbed hardware components

Generic Description	**Equipment**
Processing unit with miniPCI interface for IEEE 802.11 compatible network card connection and two USB ports for connecting peripherals	Single board computer (SBC) Alix model 3d3
Wireless network card supporting IEEE 802.11g/n	Mikrotik RouterBOARD R52Hn/R52n (dual-band) MIMO 2x2
Physical support for data storage	32GB CompactFlash card
GPS receiver	GlobalSat BU-353-S4
Accelerometer and gyroscope sensor	Sparkfun Razor IMU 9DoF
Antenna	Ubiquiti AM-5G16-120 (Shore), Ubiquiti AMO-5G10 (Ship)

The GPS receiver is integrated in the equipment installed on the vessel to enable the collection of information on the course of the ship. The ship node also includes an accelerometer and a gyroscope coupled to the antenna for the purpose of collecting data on the oscillating movement of the antenna. The information collected by the GPS, the accelerometer and the gyroscope will be used to study the impact of the distance and the antenna oscillation on the quality of the radio connection.

4.1.1.1 Accelerometer

An accelerometer is a device used to measure acceleration of motion of a structure, which is the rate of change of the velocity of an object. They sense either static or dynamic forces of acceleration. Static forces include gravity, while dynamic forces can include vibrations and movement.

Accelerometers are able to measure acceleration on one, two, or three axes. 3-axis units are becoming more common as the cost of development for them decreases.

Typically, accelerometers incorporate capacitive plates internally. Some of these are fixed, while others are attached to tiny springs that move internally as acceleration forces act upon the sensor. As these plates move in relation to each other, the capacitance between them changes. The acceleration is determined from these changes in capacitance.

The MareCom project testbed used an accelerometer with 3-axis unit.

4.1.1.2 Gyroscope

Gyroscopes, also known as gyros, are devices that measure or maintain orientation and angular velocity. Angular velocity is simply a measurement of speed of rotation around an axis. Normally, gyros are used to determine orientation and are found in most autonomous navigation systems.

As accelerometers, gyroscopes can also measure rotation around three axes: x, y and z. The triple axis gyro is becoming more popular, smaller and less expensive.

Generally, gyros are often used on objects that are not spinning very fast at all and it is important to keep in mind that acceleration or linear velocity does not affect the measurement of the gyro.

As stated in Table 4.1, the accelerometer and gyroscope sensor is the same, which implies that the gyroscope also contains 3-axis.

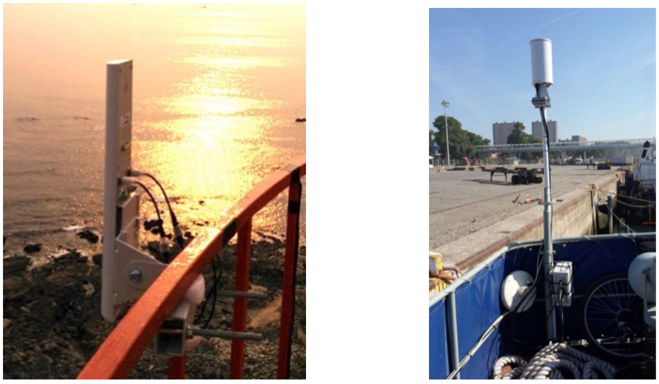
4.1.2 Software

The software component is supported by the OpenWrt [30] operating system (Linux 4.4) and intends to automate the process of collecting data on the measured signal and debit reached during the course of the vessel. The software component consists of three main modules:

- The central control module (CCM), which is present only in the BS and is responsible for monitoring the number of nodes associated with the BS, automatically triggering the test procedure always according to the effective availability of the connection with the station (STA).
- The data collection module (DCM), which is present in the BS and the STAs and is responsible for capturing and collecting various parameters during the transmission and reception of each packet, including RSSI, physical debit, packet size, antenna position relative to the vertical and geographical position of the vessel.
- The traffic generation module (TGM), which is present in the BS and the STAs and enables the generation of traffic in an automated way, which is captured by the data collection module in order to obtain a channel characterization.

4.2 Installation setup

A ground station, BS was installed at the Lighthouse of Leça, at an height of 55 meters above sea level, and two clients, stations, placed on two different vessels, at height of 10 meters above sea level, as shown in Figure 4.2.



(a) Shore antenna (b) Ship antenna

Figure 4.2: Installed antennas.

The vessel leaves at a distance of approximately 1.92 kilometers of the transmitter (the BS in the lighthouse) marked as a red line in figure 4.3.

Figure 4.3: Distance between the transmitter and the receiver at the beginning of the tests.

Both vessels were navy vessels and had the same characteristics. The length is about 48 meters, small to medium displacement (200 to 400 tons), the maximum speed is 20 knots and they are intended to operate along coastal areas in surveillance, patrol and defense missions.

During the MareCom project two communications test scenarios were defined:

- Earth-Sea Communications;
- Sea-Earth Communications,

according to the direction of the main flow.

4.3 Test weather conditions

The tests were generated on four different days at different time intervals:

- November 6, 2016, 19h13 to 22h48 (approximate duration of 4 hours)
- November 7, 2016, from 00h00 to 14h42 (approximate duration of 15h hours)
- November 16, 2016 from 12h10 to 23h54 (approximate duration of 11 hours)
- November 17, 2016 from 00h00 to 09h35 (approximate duration of 10 hours)

According to the information on the Windguru portal [31], during these days and the area where the vessel sailed, the temperature was around 13º C and didn't rain. During the first week the sea ripple were around 1.6 meters, while during the second week were around 1 meter. In terms of wind, during the first week the wind speed was around 10 knots and during the second week was around 4 knots. Summarizing, the atmospheric conditions were better during the second week.

4.4 Testbed configuration

The tests had two different approaches for the two distinct weeks. During the first week, all traffic generation was made just in one direction, where the lighthouse is the transmitter and the vessel is the receiver, being that an earth-sea communications test. During the second week, the traffic generation had two main flows, where the lighthouse may be a transmitter or a receiver, as well as the vessel, existing on this way two flows: earth-sea and sea-earth communications test.

In order to synchronize the two flows of communications, it was stipulated that in every 2 minutes the direction of communication was being changed, i.e., in the first 2 minutes the communication was made from the lighthouse to the vessel and in the following 2 minutes from the vessel to the lighthouse, until there is no signal received by the vessel.

To facilitate the understanding of coming up chapters of this dissertation, let us call the earth-sea flow as *normal* and sea-earth as *rev* as illustrated in Figure 4.4.

The servers (traffic generators), either being the vessel or the lighthouse, only generated UDP traffic and had a fixed data rate of 2 Mbit/s.

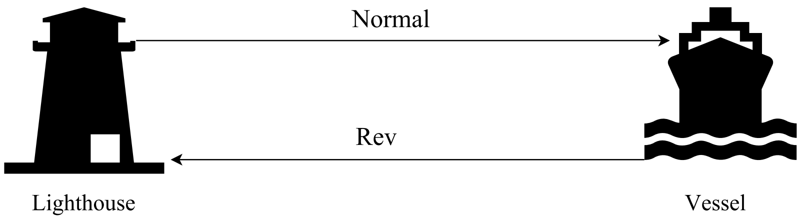


Figure 4.4: Normal and Reverse flows.

4.5 Propagation models

Before the data analysis, some investigation were made in order to understand what to expect in terms of communication range. In this section, the study of the propagation models are presented.

As stated in Chapter 2, there are two types of propagation models in maritime environment that are usually studied: free space model and 2-ray model.

Taking in account the environment where tests were developed, the transmitter and the receiver were always communicating without any obstructions nearby and we could assume that there is only one clear line-of-sight path between the transmitter and receiver, making this way the free space model a valid one.

Another approximation to the scenario where the tests were performed is illustrated in Figure 4.5. In general, there are two rays propagating between the transmitter and the receiver, a direct ray and a reflected ray on water.

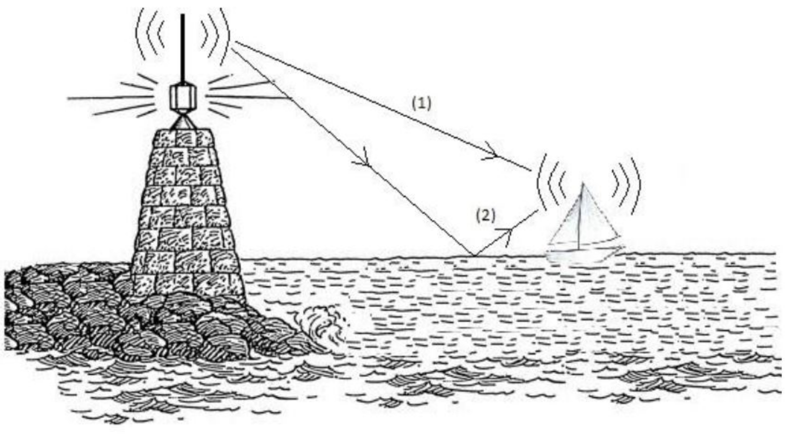


Figure 4.5: Two-ray model, where (1) is the direct ray and (2) is the reflected ray [32].

In order to determine the maximum possible distance of communication and to understand the variation of the signal as a function of distance, free space and 2-ray propagation model were studied.

Table 4.2 presents the values considered to study these propagation models.

Table 4.2: Values for propagation models

Parameter	Value
Wi-fi card output power P_t	18 dBm
Ship antenna height h_r	10 m
Shore antenna height h_t	55 m
Ship antenna gain G_r	1,5 dBi
Shore antenna gain G_t	14 dBi
Operating frequency f	500 MHz

The path loss is defined, for each model, respectively by:

Free Space Path Loss Model:

$$L_{fs} = 10\log\left(\frac{4\pi d}{\lambda}\right)^2 \tag{4.1}$$

2-Ray Path Loss Model:

$$L_{2ray} = 10\log\left(\frac{\lambda^2}{(4\pi d)^2}\left(2\sin\frac{2\pi}{\lambda}\frac{h_t h_r}{d}\right)^2\right), \tag{4.2}$$

where L is the path loss in dB, h_t and h_r are the effective heights of the transmitter t and receiver r, respectively, d is the distance between transmitter t and receiver r. The λ represents the wave length of the radio transmission in the same dimension of the antenna height and distance. In order to calculate the expected path loss for the MareCom project, it was necessary to convert the operating frequency f into wavelength λ with the following equation:

$$\lambda = \frac{c}{f}, \tag{4.3}$$

where constant c corresponds to speed of light and has a value of $3 \times 10^8 m/s$, resulting on a wavelength of:

$$\lambda = \frac{3 \times 10^8}{509 \times 10^6} = 0,589\ m. \tag{4.4}$$

Figure 4.6 represents the path loss as a function of distance for both models.

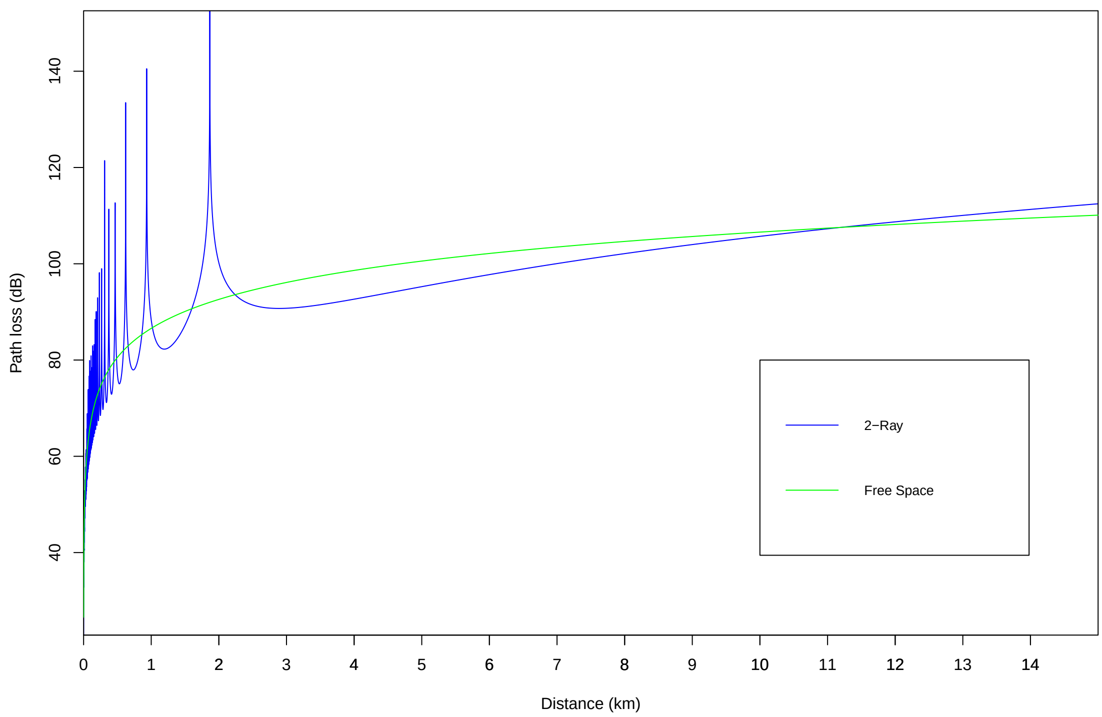


Figure 4.6: Path loss according to the 2-ray and free space path loss model.

Chapter 5

Data Processing

In this chapter, we describe the procedure adopted in order to search for and identify relationships in large amounts of information related to the transmission of data in maritime environments. Several steps were taken, such as, studying the variables presented in the logs, organizing the data, sampling the data, understanding the route made by the vessel on different days and attempting to identify the relationship between communication variables and external variables.

5.1 Logs generation

As stated in Section 4.3, the tests were performed in four different days, resulting in different logs for each day. These logs were generated by two different tools that are usually used in communications, named horst and iperf.

The second week logs (16 and 17 November) contain the two scenarios mentioned in Chapter 4, i.e., the Earth-Sea (normal) and the Sea-Earth (rev) communications scenarios, while the first week (6 and 7 November) only contains the Earth-Sea communications scenario.

5.2 Horst

The Highly Optimized Radio Scanning Tool (HORST) is a lightweight IEEE802.11 wireless LAN analyzer with a text interface [33]. Its main function is related to Wireshark, tcpdump, but it has some additional advantages, such as its small size and its ability to show particular information which is not easily available from other tools. This tool is useful to get a quick overview of what is happening on all WLAN channels and to identify problems.

5.2.1 Logs analysis

Firstly, it was necessary to study the variables presented in the logs, in order to analyze the data and to identify anomalous observations. These logs had about 40 variables, some of which did not contain values or contain constant values, making them useless for further study.

The variables that really matter for further analysis are presented in Table 5.1.

Table 5.1: Horst logs variables

Variables	Description	Units
WLAN TYPE	Package type transmitted	None
RSSI (Received Signal Strength Indicator)	Measure of power level that the receiver (vessel) is receiving from the transmitter (lighthouse) or vice versa	dB
PKT LENGTH	Size of the whole packet including the header	Bytes
NOISE	Disruption that interferes with the transmission or the interpretation of information from the sender to the receiver	dB
ACC_x, ACC_y, ACC_z	Linear acceleration on three axes	g [1]
GIR_x, GIR_y, GIR_z	Angular rotation velocity on three axes	°/s
MAG_x, MAG_y, MAG_z	Magnetic fields vectors on three axes	μT
LAT, LONG, ALT	Vessel coordinates (LAT: Latitude, LONG: Longitude, ALT: Altitude)	Degrees Decimal Minutes

The WLAN TYPE variable presented in Table 5.1 takes three different values:

- BEACON, which are management frames in IEEE 802.11 based WLANs, transmitted periodically to announce the presence of a wireless LAN;
- ACK, which are usually associated to packets related to the communications protocol;
- QDATA, which are packets containing data generated by the server.

The most common packets in the logs were QDATA. Since we are analyzing the behavior of the exchange of data packets in a maritime environment, we decided to remove the packets related to protocol (ACK) and management frames (BEACON), maintaining only the QDATA packets, which were similar packets in terms of size.

[1] A single G-force on planet Earth is equivalent to 9.8 m/s^2.

5.2.2 Logs gathering

Each test day produced a large amount of horst files, around one hundred per day, being each file composed by approximately 27 thousand lines. For gather all the logs of a day into one data set in a controlled and sequential way, a script in R was developed. Figure 5.1 illustrates this script structure.

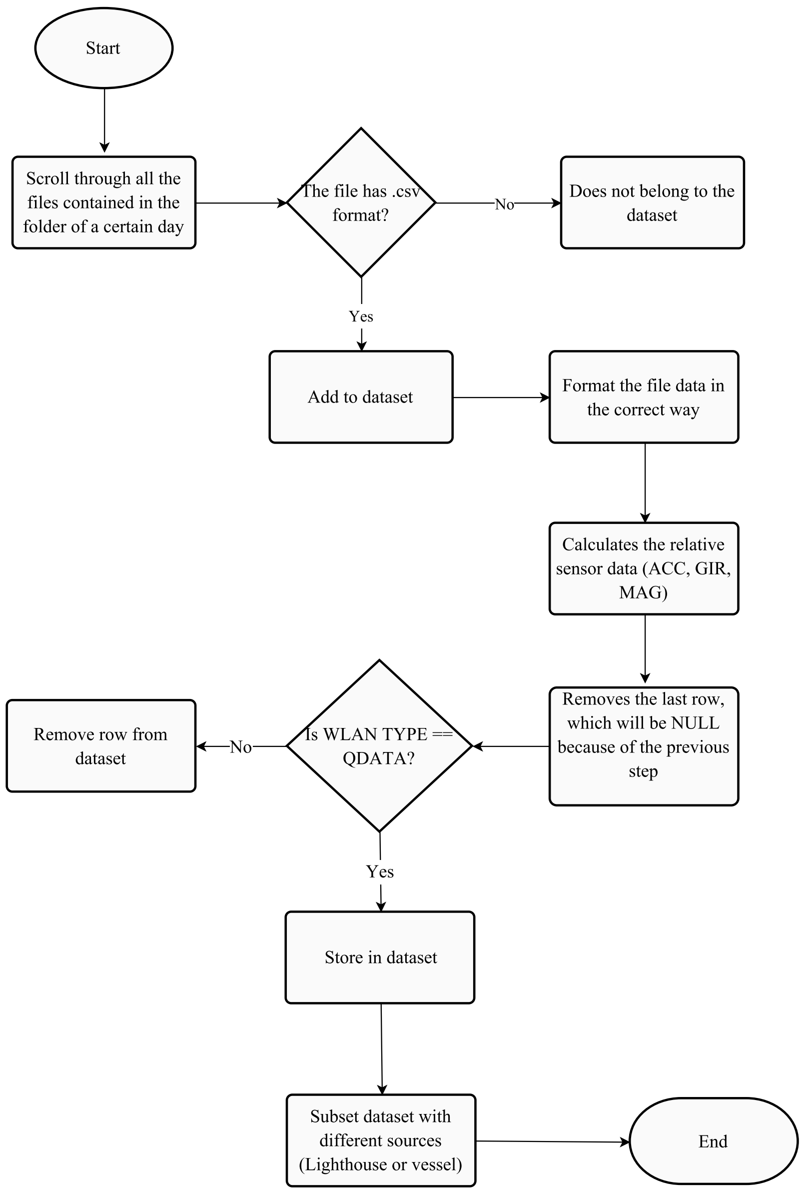


Figure 5.1: Script flow chart for data set creation.

Initially, the logs produced were not formatted correctly, since the values of some columns did not belong to the respective header. Thus, for each file, it was necessary to remove the header and format it correctly, placing the column values matched with the correct header.

Since during the tests the accelerometer, gyroscope and magnetometer sensor was not calibrated, in order to carry out a more detailed analysis, it was decided to calculate the relative values of that sensor, i.e., to calculate the difference between the current and the previous value of the accelerometer, the gyroscope and the magnetometer.

After the relative data of the sensor has been calculated, the 3-axis units of the acceleration were converted to absolute value using the following equation:

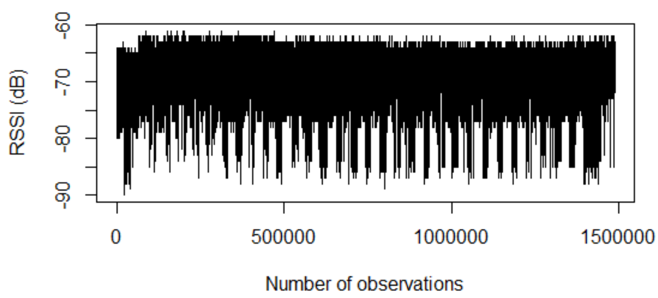
$$ACC = \sqrt{ACC_x^2 + ACC_y^2 + ACC_z^2}. \tag{5.1}$$

The other sensor values (gyroscope and magnetometer) will be covered in the following sections.

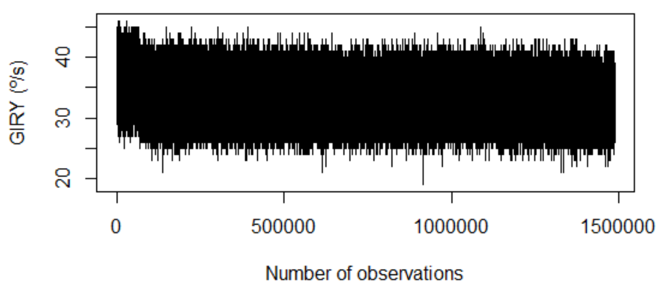
Each test day presented data logs with some format variations, leading to the development of a script for each day. However, all the scripts were based on the same algorithm represented on Figure 5.1.

The data in the gathered file is sequential in time and have millions of observations, making it difficult to draw any conclusions. Plots were made, in order to make it easier to identify anomalous situations.

Figures 5.2 and 5.3 shows examples of how the data was initially unsampled, with the source being the BS in the lighthouse.

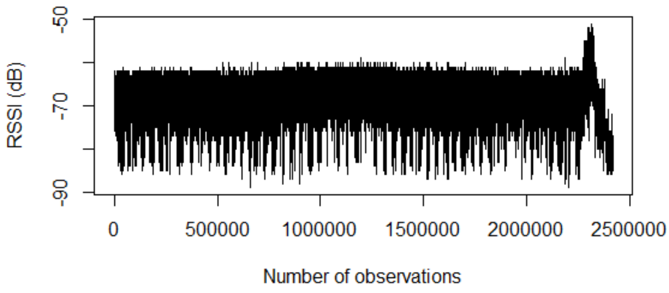


(a)

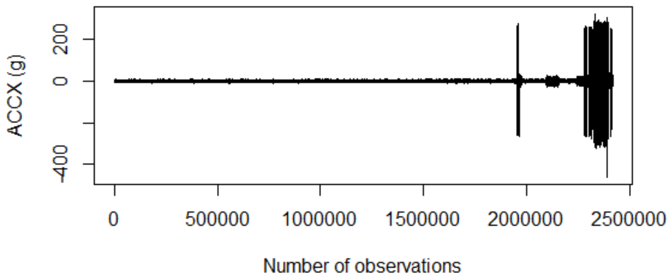


(b)

Figure 5.2: (a) RSSI and (b) gyroscope in Y-axis values on November 16.



(a)



(b)

Figure 5.3: (a) RSSI and (b) accelerometer in X-axis values on November 17.

5.2.3 Data sampling

As stated in the previous section, it is almost impossible to draw any conclusion about some aspects of the data due to the large number of observations. This led to two questions concerning data sampling: 1. which is the best sampling method to apply and 2. what should be the best sampling factor.

To overcome the first question, two methods of data sampling were attempted. The first method selects every *n*th row for a data set, where *n* is the sampling factor, and save that value. For example, one possibility would be to select every 4th row producing a data set that is 25% of the original size. The second method, which was advised by MareCom project developers, samples the data using a means approach. Basically, it selects *n* observations of the data set, calculates the mean value and saves it. This procedure is done until the end of the data set. After comparing the two methods, the one which had a better performance was undoubtedly the sample means, being on this way the one applied in this work.

In relation to the sampling factor, we took as criteria the iperf logs. Since the location of the vessel is not indicated in the iperf logs, the way we found to relate the vessel location and those logs was to sample the horst files to have the same number of iperf observations, making on this way possible to relate the location of the vessel with the variables presented in the iperf logs, which were the bandwidth, the jitter and the percentage of lost packets.

5.2.4 Vessel route

After the data set for each day have been generated and sampled, a script in R was made to convert the coordinates variables from Decimal Degree Minute (DDM) to Decimal Degree (DD) and to calculate the distance between the vessel's current point and the transmitter in the lighthouse. After the previous step is done, another script was developed using python, which produces a KML file used to represent the route of the vessel for the different days in Google Earth.

(a)

(b)

(c)

Figure 5.4: Route made by vessel on November (a) 6, (b) 7 and (c) 16.

5.2.5 Data selection

As it is possible to observe in Figure 5.4 and 5.5, only on November 17 the Wi-Fi study in the maritime environment was actually put into practice. This leads us to have less data to be analyzed, since the four days of tests were reduced by only one.

As stated in Section 4.3, on November 17 the tests were carried out for about 10 hours and vessel reached 13,5 kilometers from the lighthouse, which led us to conclude that the vessel was not always on the move. Taking into account this situation, it was concluded that the ideal procedure for the data analytics would be to focus on the November 17 and divide the data into two different scenarios: a static scenario, where the vessel is anchored and a moving scenario, where the vessel is moving. The static scenario time interval was between 00h00 and 08h50, while the moving scenario time interval was from 08h50 to 09h35.

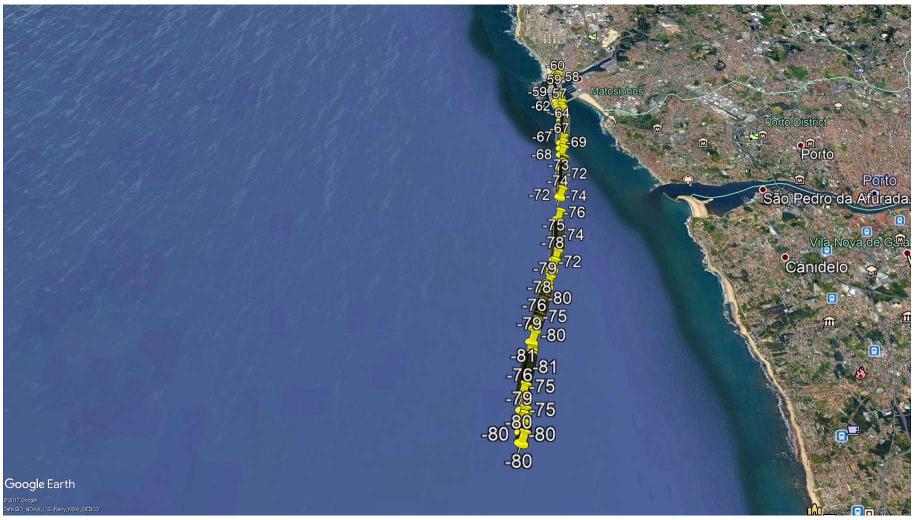


Figure 5.5: Route made by vessel on November 17.

5.3 Iperf

Iperf is a widely-used tool that was developed for measuring TCP and UDP bandwidth performance [34]. Iperf has client and server functionalities, and is able to create data streams to measure the throughput between the two ends in one or both directions. The user is able to enter iperf commands that will produce a vision on the network's delay, bandwidth availability, jitter and data loss, by tuning several parameters and characteristics of the TCP/UDP protocol.

5.3.1 Logs analysis

The same procedure used in horst was made for iperf, but these logs only contain five variables that are displayed in Table 5.2.

Table 5.2: Iperf logs variables

Variables	**Description**	**Units**
Interval	Interval of time between measurements. The time interval is set by the user.	s
Transfer	The amount of data transmitted during a time interval.	Bytes
Bandwidth	The link speed registered during a time interval.	Bit/s
Jitter	The latency variation for a given time interval.	ms

5.3.2 Logs gathering

Such as horst, many iperf files were generated during the different days. A script was developed to gather them for later analysis. As stated in Section 5.2.3, since the data sampling were made taking in consideration the number of iperf observations, it was possible to relate the iperf data sets with the distance to the lighthouse.

Chapter 6

Data Analytics

This chapter focuses on understanding various aspects of maritime communications, such as, the impact of the vessel movement in the communication, the influence of some variables on the RSSI, and finally which propagation model is best suited for maritime communications.

As stated in Section 4.4, there are two types of communication flows considered in this study: normal and rev. Since the RSSI is only measured in the vessel, all data generated in the rev direction will have a 0 value for the RSSI, except for the protocol packets (ACK) that will be sent from the transmitter in the lighthouse to the vessel's receiver to establish the communication. Also, as referred in Section 5.2.1, the packets containing data (QDATA) were kept in the data set and the protocol packets have been removed. Taking all this into account, it was decided that data analytics would be done only in the normal flow, where all packets are QDATA, which have similar length and have measured RSSI values.

6.1 Movement impact

Before the statistical models were applied, a study were made focused on identifying relationships between data communication variables and external variables, such as environmental variables, direction of the vessel, location, etc. As stated in Section 5.2.4, we will compare two different scenarios, one for the vessel in motion and the other for a static vessel.

6.1.1 Static scenario

In this section, the objective is to identify external variables that may influence the communication variables while the vessel is anchored and then compare with the scenario in which the vessel is moving away from the shore. Even with the vessel in a static scenario, although the distance does not increase, there is always sea oscillation, which may or may not affect the quality of the connection.

Figures 6.1, 6.2 and 6.3 illustrate some communication and external variables as a function of time, while the vessel is anchored.

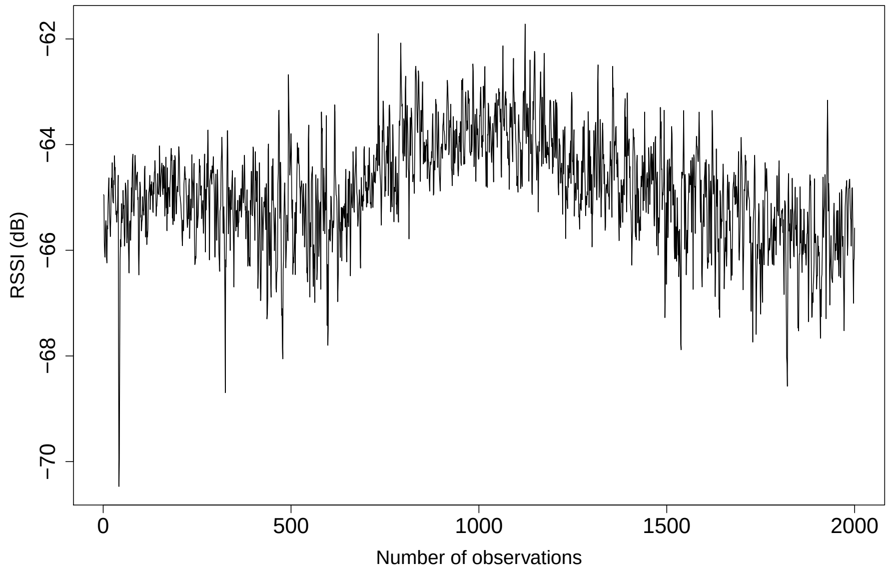


Figure 6.1: RSSI while the vessel is anchored.

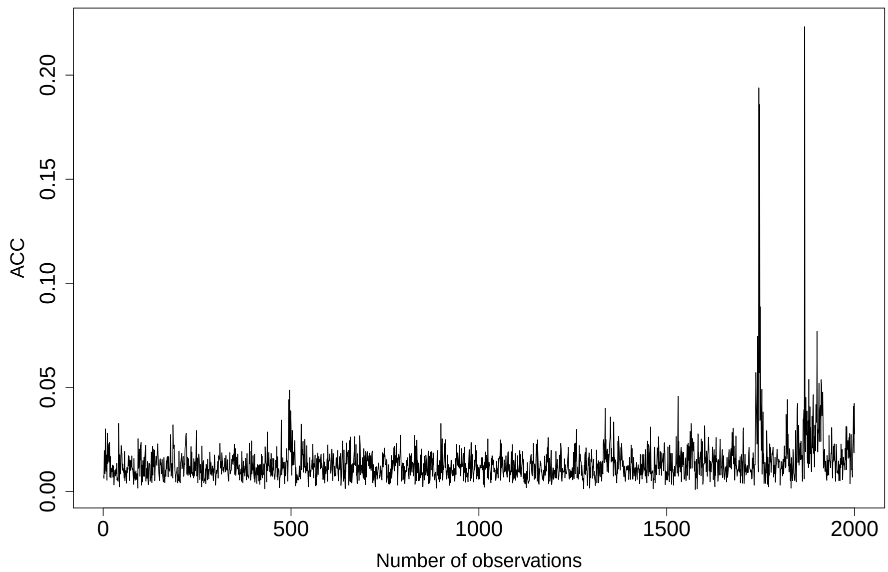


Figure 6.2: Accelerometer values while the vessel is anchored.

By analyzing these figures, it is possible to understand that while the vessel is anchored, there are no major variation in terms of power received (RSSI), antenna acceleration and noise. This leads us to conclude that while the boat is anchored, external factors have little or no effect on the quality of the connection.

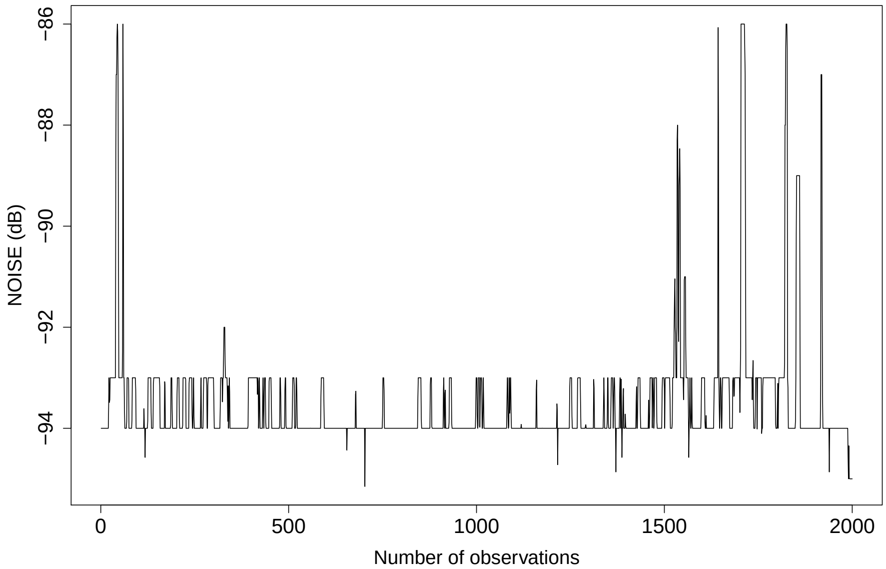


Figure 6.3: Noise while the vessel is anchored.

6.1.2 Moving scenario

Contrary to the previous scenario, in this section we want to demonstrate that while the vessel is moving, external variables will produce a greater impact on communication variables.

Figure 6.4 illustrates the RSSI value as a function of time, while the vessel is moving away from the shore.

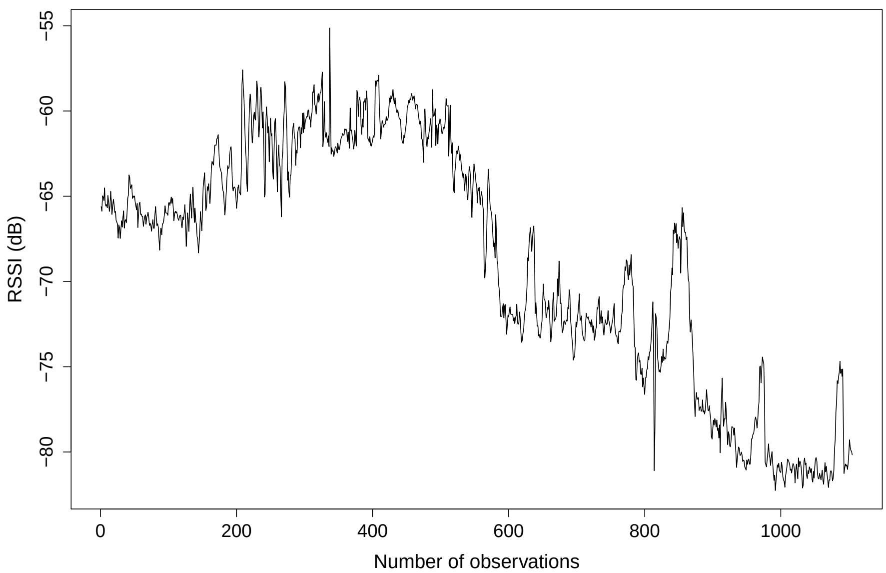


Figure 6.4: RSSI while the vessel is moving away from the shore.

Analyzing Figure 6.4, it is possible to notice that the power received by the vessel has a completely different behavior from the one shown in Figure 6.1. This scenario was already predictable, since the power received by the receiver, regardless of the environment, always has a decrease with the increase of the distance between the transmitter and the receiver.

In relation to the accelerometer, presented in Figure 6.5, before the number of observation is 400, the acceleration seems to be quite stable, but after this number of observations it is possible to detect a very significant growth. This spontaneous growth is present in all referential components of the accelerometer, as for example in Figure 5.3. Initially, we thought it was an anomal situation, but after a further analysis, this growth had an explanation given by the environment where the vessel is at that time, which is the open sea.

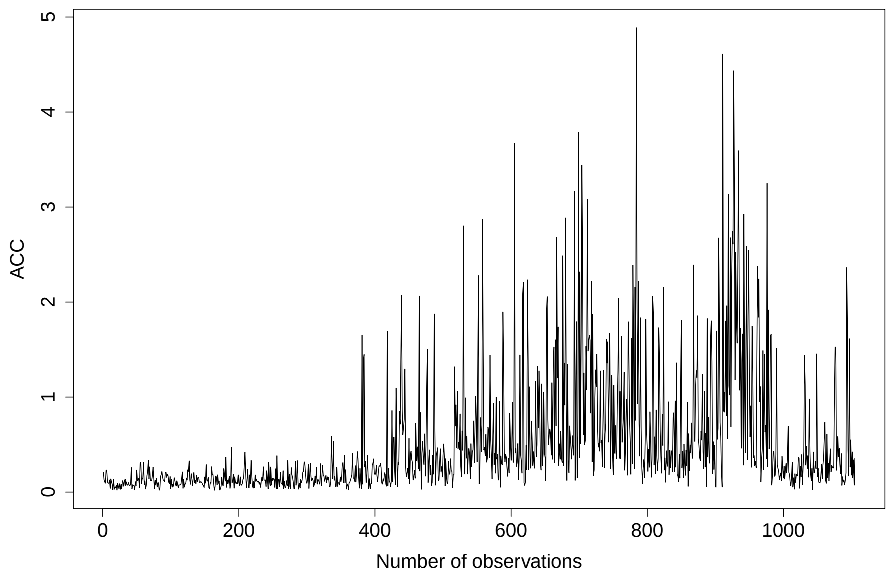


Figure 6.5: Accelerometer values while the vessel is moving away from the shore.

Figure 6.6 represents the accelerometer values in function of distance. It is possible to verify that the acceleration increases when the vessel is up to 4 kilometers from the shore. The gaps with no registered values, correspond to the rev communication flow, which was not included in this study, as mentioned at the beginning of the chapter.

With the help of Google Earth and taking into account the generated data with normal flow, we can see that the vessel is no longer protected by the harbor, when it is 4 kilometers away from the shore, as shown in the red line in Figure 6.7. This explains the considerable increase in the accelerometer values, since the vessel suffers a greater sea oscillation.

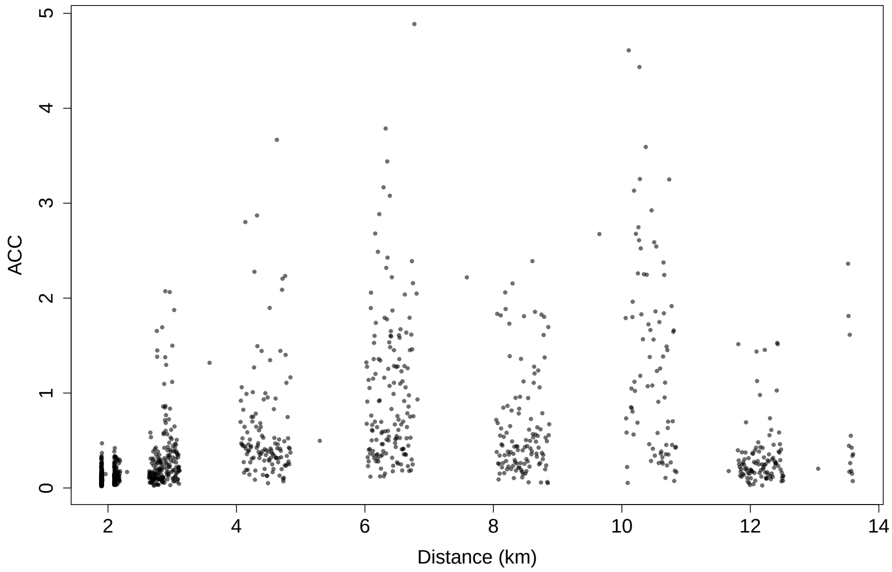


Figure 6.6: Accelerometer values as function of distance.

Figure 6.7: Vessel 4 kilometers away from the shore.

In relation to noise, shown in Figure 6.8, it is possible to verify that it takes higher values as the vessel moves away from the shore, which does not happen during the static scenario, where the noise remains practically constant except for some values that may have been triggered by some kind of interference.

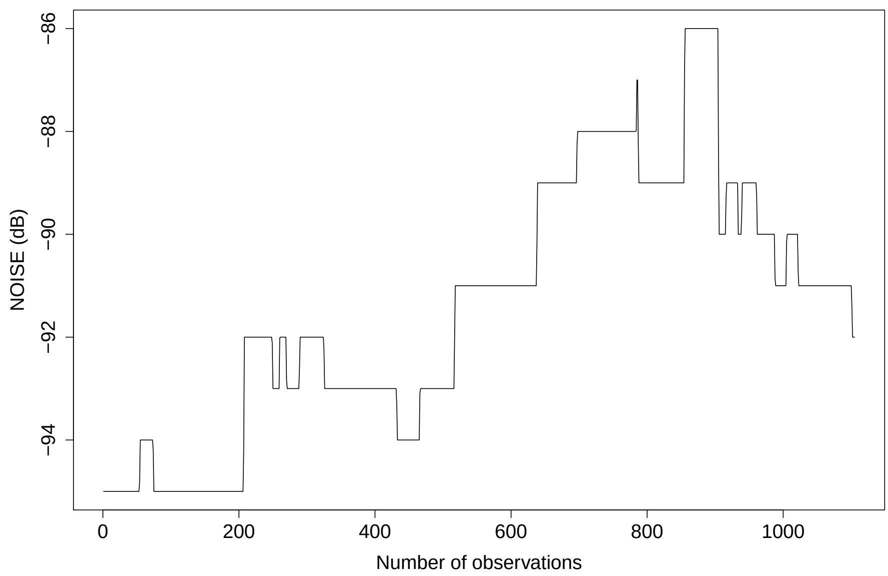


Figure 6.8: Noise values while the vessel is moving away from the shore.

As stated in Section 4.4, the data rate was fixed at 2 Mbits/s, which leads to not being able to draw great conclusions about the variables present in the iperf logs.

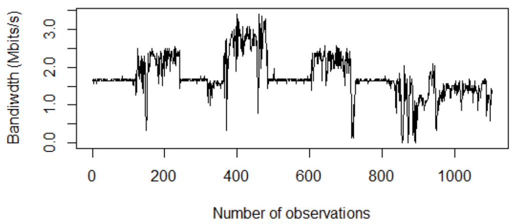


Figure 6.9: Bandwidth while the vessel is moving away from the shore.

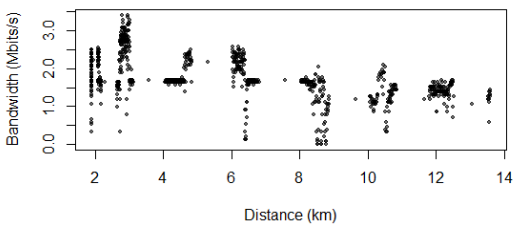


Figure 6.10: Bandwidth as function of distance

In Figures 6.9 and 6.10, we can observe the plots related to the bandwidth, one in function of the time and the other in function of the distance, respectively. It is possible to verify that although the bandwidth attempts to establish in the 2 Mbits/s, as the vessel moves away of the shore the decreases in bandwidth become more recurrent. This situation can happen due to the fact that the sea oscillation constantly changes the position of the boat and consequently the boat antenna loses connectivity with the ground antenna, which may result in a temporary loss of signal during the generation of traffic.

Figures 6.11 and 6.12 illustrate the plots related to the jitter, one in function of time and the other in function of distance, respectively. It is verified that while the vessel is near the shore, there is a tendency for a stable connection and while the vessel is moving away the values be-

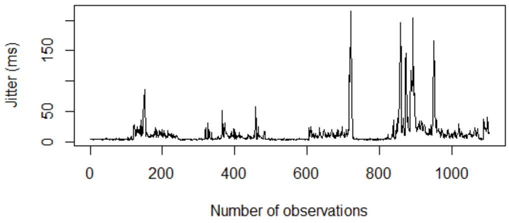


Figure 6.11: Jitter while the vessel is moving away from the shore.

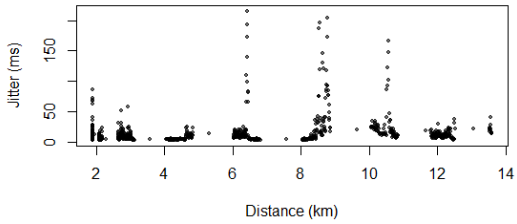


Figure 6.12: Jitter as function of distance

come higher, which indicates that re transmissions might be occurring due to Wi-Fi connection instability.

6.2 Regression analysis

In this section, regression will be applied in order to estimate the relationship among the identified variables.

6.2.1 Correlation matrices

In order to better understand the relationship between the variables, the correlation matrices of the variables of interest for static and moving scenarios were made for November 17 with the source being the lighthouse, and are presented in Table 6.1 and 6.2.

Table 6.1: Static scenario correlation matrix.

	RSSI	NOISE	PKT_LENGTH	ACCX	ACCY	ACCZ	GIRX	GIRY	GIRZ	MAGX	MAGY	MAGZ	LAT	LONG	ALT	ACC
RSSI	1.000	-0.188	0.003	0.021	0.005	0.015	-0.013	-0.024	0.013	-0.040	-0.023	-0.006	-0.028	-0.155	0.300	-0.196
NOISE	-0.188	1.000	-0.028	-0.002	-0.021	-0.019	0.007	0.020	0.055	0.017	0.003	-0.020	0.001	0.005	-0.209	0.004
PKT_LENGTH	0.003	-0.028	1.000	0.035	-0.017	0.004	0.002	-0.034	0.003	0.026	-0.004	-0.033	-0.008	-0.001	0.026	-0.041
ACCX	0.021	-0.002	0.035	1.000	0.031	0.111	-0.017	0.040	-0.020	-0.025	0.022	-0.018	-0.027	-0.020	0.036	0.103
ACCY	0.005	-0.021	-0.017	0.031	1.000	0.118	-0.087	-0.002	-0.006	-0.077	-0.193	-0.045	-0.044	-0.008	0.040	-0.191
ACCZ	0.015	-0.019	0.004	0.111	0.118	1.000	0.022	0.015	-0.013	-0.014	-0.007	-0.029	0.011	-0.007	0.000	-0.048
GIRX	-0.013	0.007	0.002	-0.017	-0.087	0.022	1.000	0.012	0.004	-0.003	0.032	0.039	0.003	-0.005	-0.004	0.003
GIRY	-0.024	0.020	-0.034	0.040	-0.002	0.015	0.012	1.000	0.002	-0.043	-0.013	0.013	-0.012	0.029	-0.003	0.042
GIRZ	0.013	0.055	0.003	-0.020	-0.006	-0.013	0.004	0.002	1.000	0.011	0.008	0.031	-0.007	0.020	-0.039	0.013
MAGX	-0.040	0.017	0.026	-0.025	-0.077	-0.014	-0.003	-0.043	0.011	1.000	0.239	0.002	0.006	0.004	0.000	0.002
MAGY	-0.023	0.003	-0.004	0.022	-0.193	-0.007	0.032	-0.013	0.008	0.239	1.000	0.166	0.019	0.009	0.017	-0.007
MAGZ	-0.006	-0.020	-0.033	-0.018	-0.045	-0.029	0.039	0.013	0.031	0.002	0.166	1.000	0.016	-0.006	-0.014	-0.025
LAT	-0.028	0.001	-0.008	-0.027	-0.044	0.011	0.003	-0.012	-0.007	0.006	0.019	0.016	1.000	0.265	-0.181	0.021
LONG	-0.155	0.005	-0.001	-0.020	-0.008	-0.007	-0.005	0.029	0.020	0.004	0.009	-0.006	0.265	1.000	-0.208	0.130
ALT	0.300	-0.209	0.026	0.036	0.040	0.000	-0.004	-0.003	-0.039	0.000	0.017	-0.014	-0.181	-0.208	1.000	-0.123
ACC	-0.196	0.004	-0.041	0.103	-0.191	-0.048	0.003	0.042	0.013	0.002	-0.007	-0.025	0.021	0.130	-0.123	1.000

Analyzing both correlation matrices, it is possible to notice that there are some variables that have a considerable correlation with RSSI, such as: noise, acc, lat, long, alt and distance. Comparing Table 6.1 and 6.2, it is possible to detect that while the vessel is moving, the correlation of some variables increases or decreases and even in particular variables, they have the opposite influence on the RSSI.

Taking in consideration the correlation matrices and the type of antennas used during the project, it is possible to verify that the gyroscope values will not have a great impact on the RSSI.

Table 6.2: Moving scenario correlation matrix.

	RSSI	NOISE	PKT_LENGTH	ACCX	ACCY	ACCZ	GIRX	GIRY	GIRZ	MAGX	MAGY	MAGZ	LAT	LONG	ALT	ACC	DISTANCE
RSSI	1.000	-0.606	-0.033	0.001	0.024	-0.005	-0.040	0.013	-0.018	0.086	-0.023	-0.007	0.904	-0.837	-0.327	-0.340	-0.905
NOISE	-0.606	1.000	0.049	-0.018	-0.045	0.028	0.028	-0.040	0.050	-0.027	0.036	-0.010	-0.679	0.432	0.200	0.449	0.675
PKT_LENGTH	-0.033	0.049	1.000	-0.014	0.009	0.019	-0.007	-0.032	0.040	0.008	0.002	0.028	-0.027	0.011	0.097	-0.023	0.026
ACCX	0.001	-0.018	-0.014	1.000	0.239	-0.416	0.081	0.584	-0.455	-0.037	-0.087	-0.032	-0.006	0.018	-0.001	-0.081	0.007
ACCY	0.024	-0.045	0.009	0.239	1.000	-0.208	-0.373	0.162	0.090	-0.042	-0.105	-0.014	0.018	0.001	-0.002	-0.197	-0.018
ACCZ	-0.005	0.028	0.019	-0.416	-0.208	1.000	-0.088	-0.567	0.372	-0.099	0.115	0.005	-0.005	-0.005	0.000	0.029	0.005
GIRX	-0.040	0.028	-0.007	0.081	-0.373	-0.088	1.000	0.064	-0.413	0.043	0.072	0.042	-0.032	0.022	0.050	0.122	0.031
GIRY	0.013	-0.040	-0.032	0.584	0.162	-0.567	0.064	1.000	-0.564	-0.067	-0.084	-0.062	0.010	0.003	0.003	-0.038	-0.010
GIRZ	-0.018	0.050	0.040	-0.455	0.090	0.372	-0.413	-0.564	1.000	0.061	0.065	0.070	-0.018	0.006	-0.010	0.000	0.018
MAGX	0.086	-0.027	0.008	-0.037	-0.042	-0.099	0.043	-0.067	0.061	1.000	0.486	0.428	0.070	-0.065	0.058	-0.024	-0.069
MAGY	-0.023	0.036	0.002	-0.087	-0.105	0.115	0.072	-0.084	0.065	0.486	1.000	0.455	-0.030	0.022	-0.044	0.062	0.029
MAGZ	-0.007	-0.010	0.028	-0.032	-0.014	0.005	0.042	-0.062	0.070	0.428	0.455	1.000	-0.013	0.018	-0.021	0.013	0.013
LAT	0.904	-0.679	-0.027	-0.006	0.018	-0.005	-0.032	0.010	-0.018	0.070	-0.030	-0.013	1.000	-0.946	-0.328	-0.335	-1.000
LONG	-0.837	0.432	0.011	0.018	0.001	-0.005	0.022	0.003	0.006	-0.065	0.022	0.018	-0.946	1.000	0.323	0.214	0.948
ALT	-0.327	0.200	0.097	-0.001	-0.002	0.000	0.050	0.003	-0.010	0.058	-0.044	-0.021	-0.328	0.323	1.000	0.124	0.329
ACC	-0.340	0.449	-0.023	-0.081	-0.197	0.029	0.122	-0.038	0.000	-0.024	0.062	0.013	-0.335	0.214	0.124	1.000	0.333
DISTANCE	-0.905	0.675	0.026	0.007	-0.018	0.005	0.031	-0.010	0.018	-0.069	0.029	0.013	-1.000	0.948	0.329	0.333	1.000

One explanation for this situation might be the wide aperture in the vertical axis of the radiation pattern of the antennas used, that even under conditions of considerable oscillation, the angular rotation velocity is not enough to degrade transmitted/received power.

Regarding the magnetometer values recorded by the sensor, there is no relationship with Wi-Fi communications and the magnetic fields that can be considered for this project.

In Figure 6.13, a scatterplot is illustrated, which shows in greater detail the relation of the RSSI and the three predictors that will be used in this section which are the distance, the antenna acceleration and the noise. The first column shows the relationships between RSSI and each of the predictors. The scatterplots show negative relationships with noise, distance and vessel antenna's acceleration. The strength of these relationships are shown by the correlation coefficients across the first row. The remaining scatterplots and correlation coefficients show the relationships between the predictors.

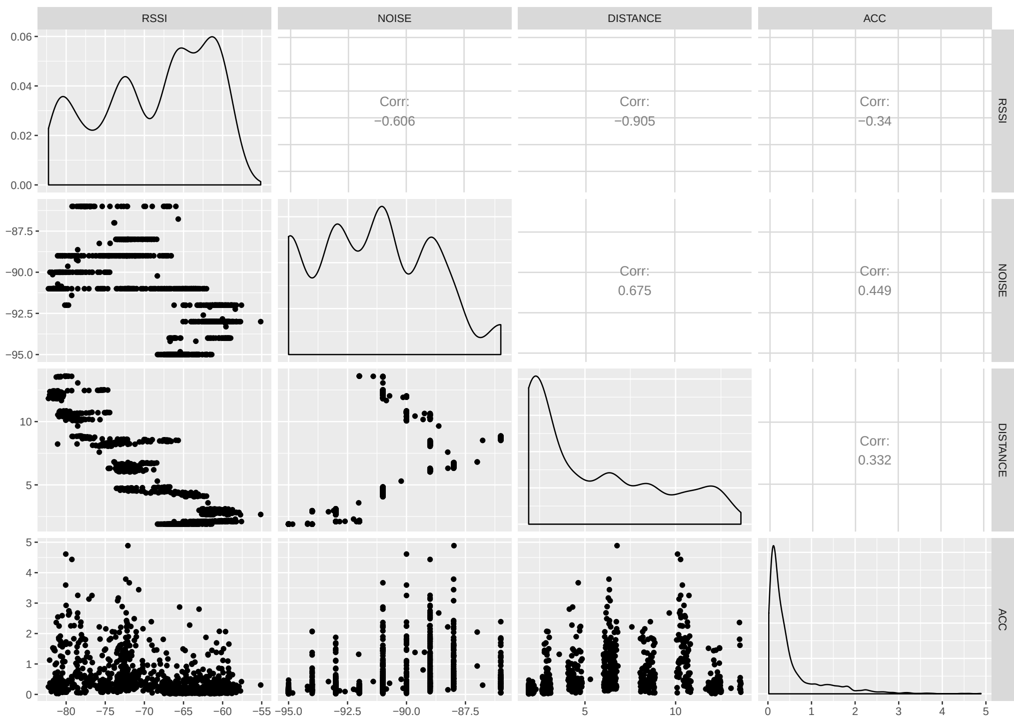


Figure 6.13: A scatterplot matrix of RSSI and the three selected predictors.

6.2.2 Simple regression

During this section, simple regression was applied, where the response variable Y is the RSSI and the predictor variables are the distance, the antenna's acceleration and noise.

1. *Simple regression where RSSI is the response variable and the distance is the predictor variable.*

   In Figure 6.14, the regression line represented in red is given by

$$RSSI = \beta_0 + \beta_1 Distance, \tag{6.1}$$

   where $\beta_0 = -59.02$ and $\beta_1 = -1.80$. Taking this values in consideration, it is possible to conclude that when the vessel is in the initial position (1.92 kilometers from the transmitter antenna) the power received is -59.02 dB and that as the vessel moves away 1 kilometer the power will decrease by 1.80 dB.

   As stated in Section 3.5, another way to illustrate applied regression is through regression trees. In Figure 6.15, the regression tree where RSSI is the response variable and the distance is the predictor variable is illustrated. It confirms that the greater the distance the smaller the RSSI, although in short distances (< 4.5 km) this is not so pronounced, because a reliable communication is still possible and the boat is not yet in a high oscillation environment, as is shown in Section 6.1.2. The percentage values represent the number of observations for a given distance.

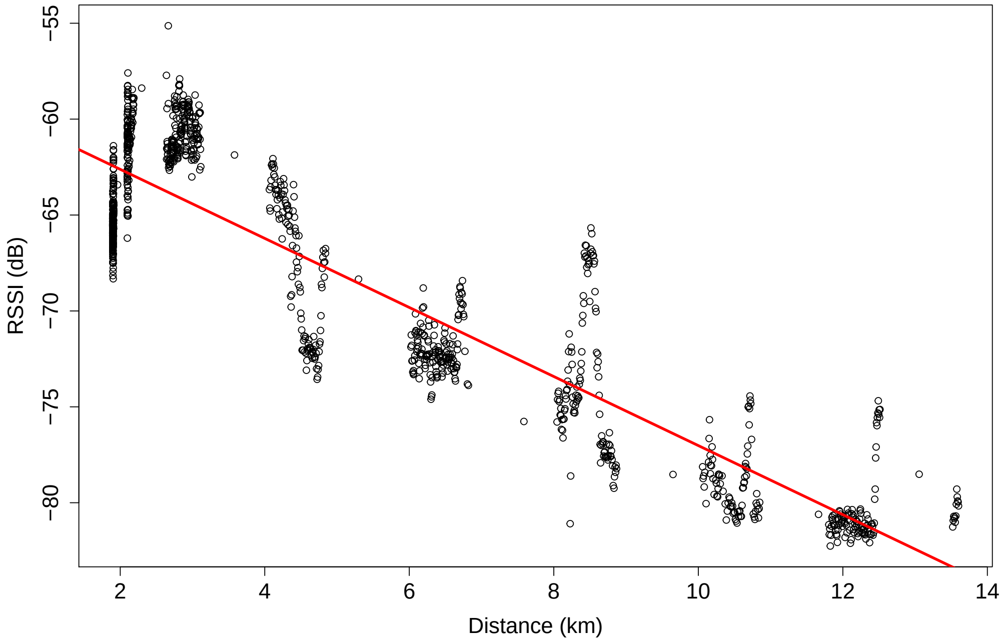


Figure 6.14: RSSI as a function of distance.

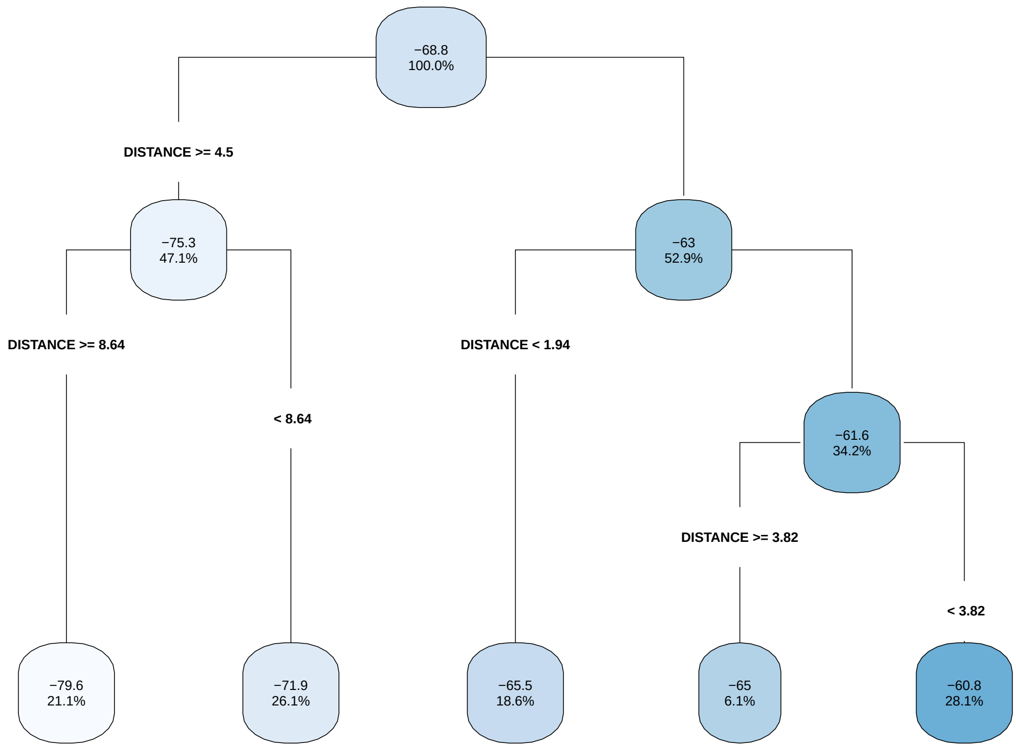


Figure 6.15: Regression tree: RSSI (dB) as function of distance (km).

2. *Simple regression where RSSI is the response variable and the vessel antenna's acceleration is the predictor variable.*

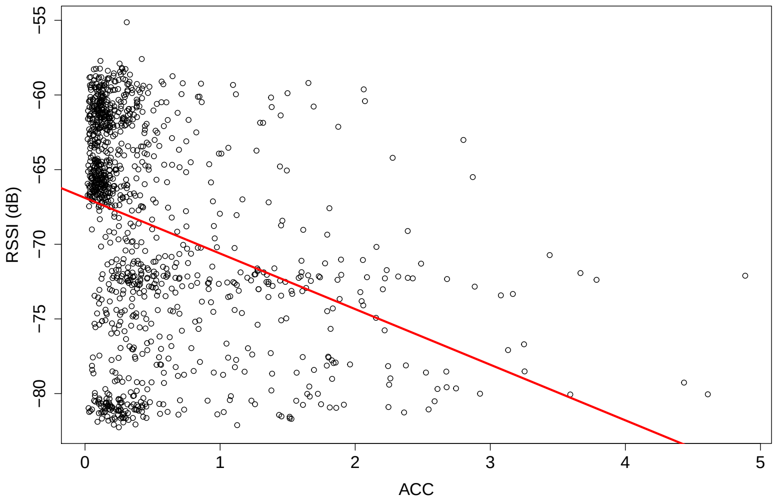


Figure 6.16: RSSI as a function of vessel's antenna acceleration.

The regression line represented in Figure 6.16 corresponds to the following model:

$$RSSI = \beta_0 + \beta_1 ACC, \tag{6.2}$$

where $\beta_0 = -66.895$, which means that in the absence of acceleration in vessel's antenna, the power received is -66.895 while $\beta_1 = -3.72$ meaning that a unit increase on acceleration of vessel's antenna will produce a reduction of 3.72 dB in the power received.

Figure 6.17 illustrates the corresponding regression tree. It is possible to easily verify that the sea oscillation influences the RSSI. If we go through the regression tree from left to right, we observe that while accelerometer values are increasing, the RSSI is decreasing.

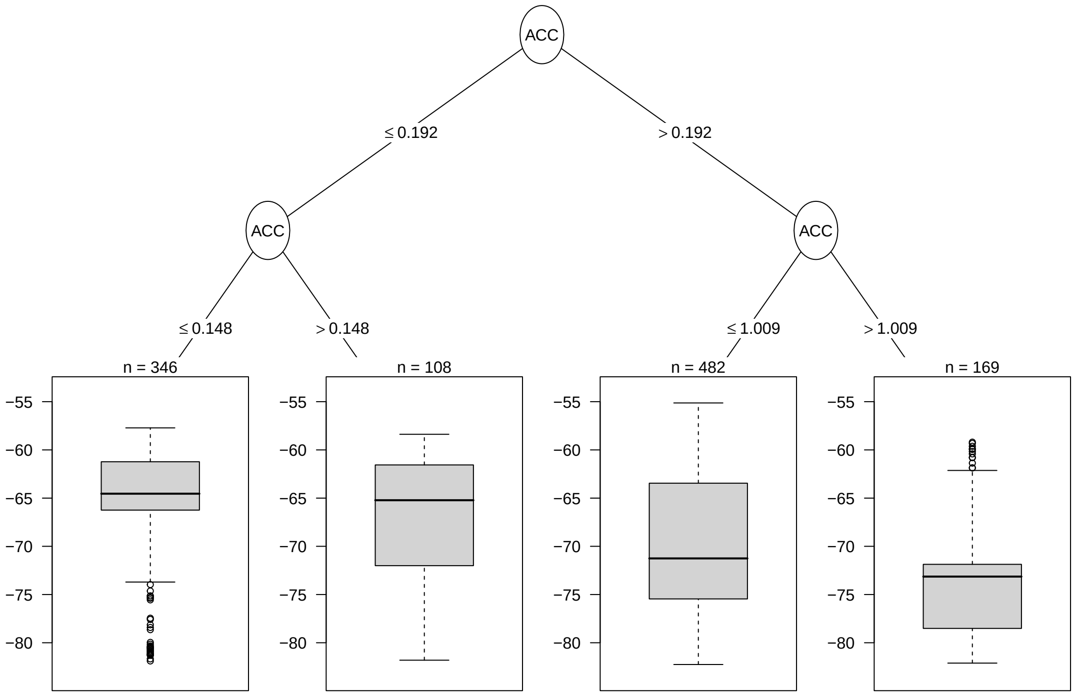


Figure 6.17: Regression tree: RSSI (dB) as function of vessel's antenna acceleration.

3. *Simple regression where RSSI is the response variable and the channel noise is the predictor variable.*

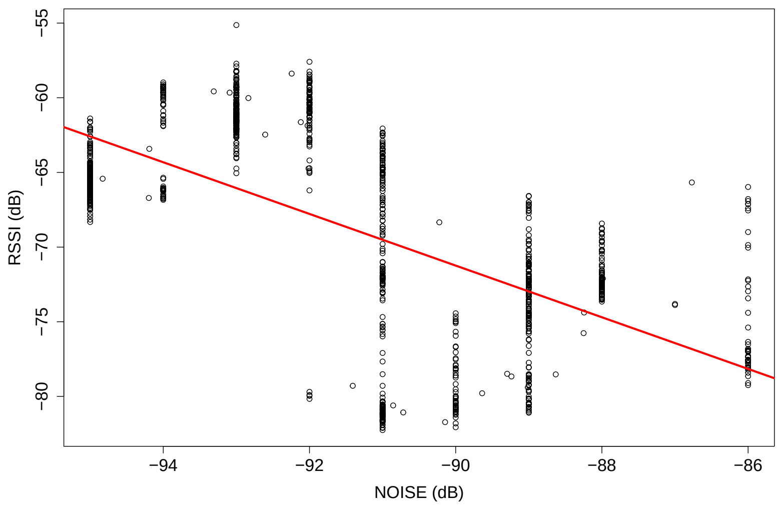


Figure 6.18: RSSI as a function of noise.

In Figure 6.18, the regression line represented in red is given by

$$RSSI = \beta_0 + \beta_1 NOISE, \tag{6.3}$$

where, $\beta_0 = -226.96$ and $\beta_1 = -1.73$. β_1 value means that if the noise increase in 1 dB, the RSSI will decrease 1.73 dB. β_0 in this case does not have a realistic meaning, because a 0 dB value for the noise does not occur.

In Figure 6.19, the regression tree where RSSI is the response variable and the channel noise is the predictor variable is illustrated.

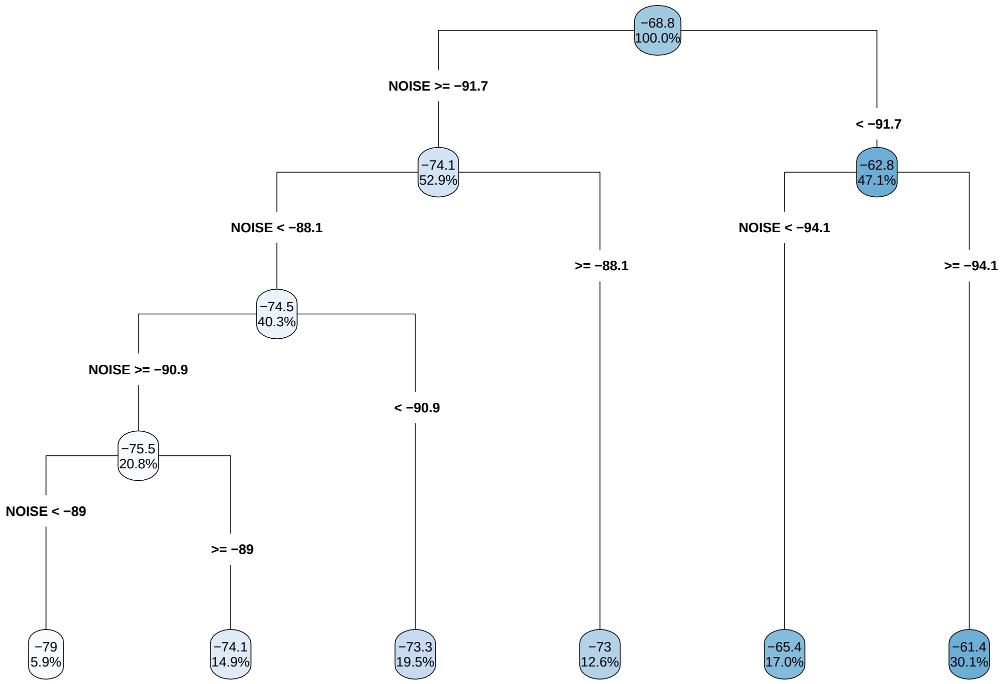
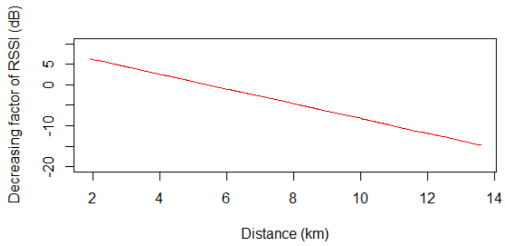
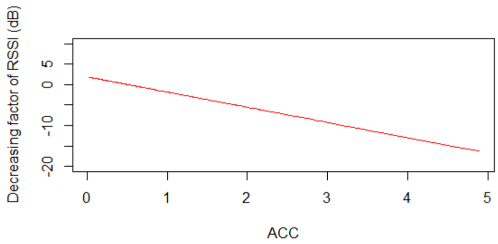


Figure 6.19: Regression tree: RSSI (dB) as function of channel noise (dB).

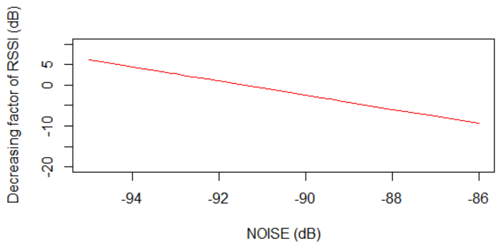
Figure 6.20 illustrates the decreasing factor of RSSI as function of the three predictor variables.



(a)



(b)



(c)

Figure 6.20: RSSI decreasing factor as function of (a) distance, (b) vessel's antenna acceleration and (c) noise

6.2.3 Multiple regression

1. *Multiple regression where RSSI is the response variable and the distance and noise are the predictor variables.*

   Using three particular values of noise, it was produced a multiple regression model that makes a prediction of what would be the behavior of RSSI in relation to distance. Basically, a multiple regression model was fitted and the predicted lines were plotted for certain levels of one predictor variable (noise).

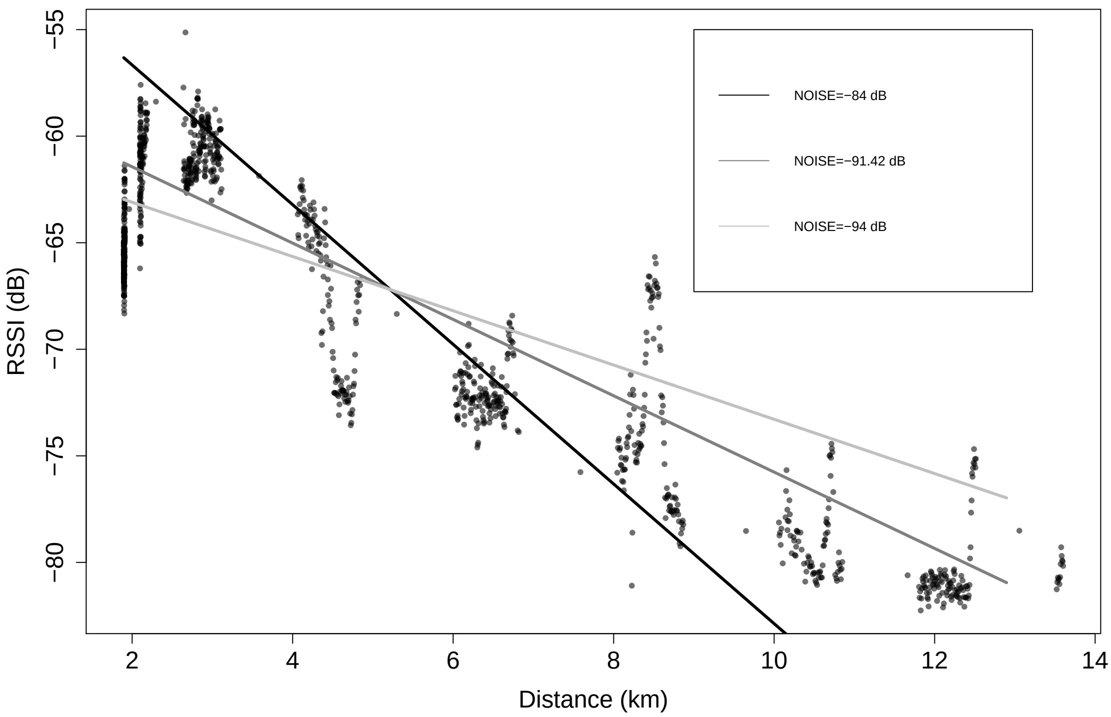


Figure 6.21: RSSI as function of distance and predictions of different noise values.

   Taking into account Figure 6.21, it is possible to conclude that for higher values of noise, higher line slope values were obtained. There is an interaction between noise and distance and we can observe that the relationship between RSSI and distance differs on the levels of noise, which means that noise has influence in this relationship.

2. *Multiple regression where RSSI is the response variable and the distance and vessel's antenna acceleration values are the predictor variables.*

   Using the same approach, a multiple regression model was fitted and the predicted lines were plotted for certain levels of one predictor variable (vessel antenna's acceleration).

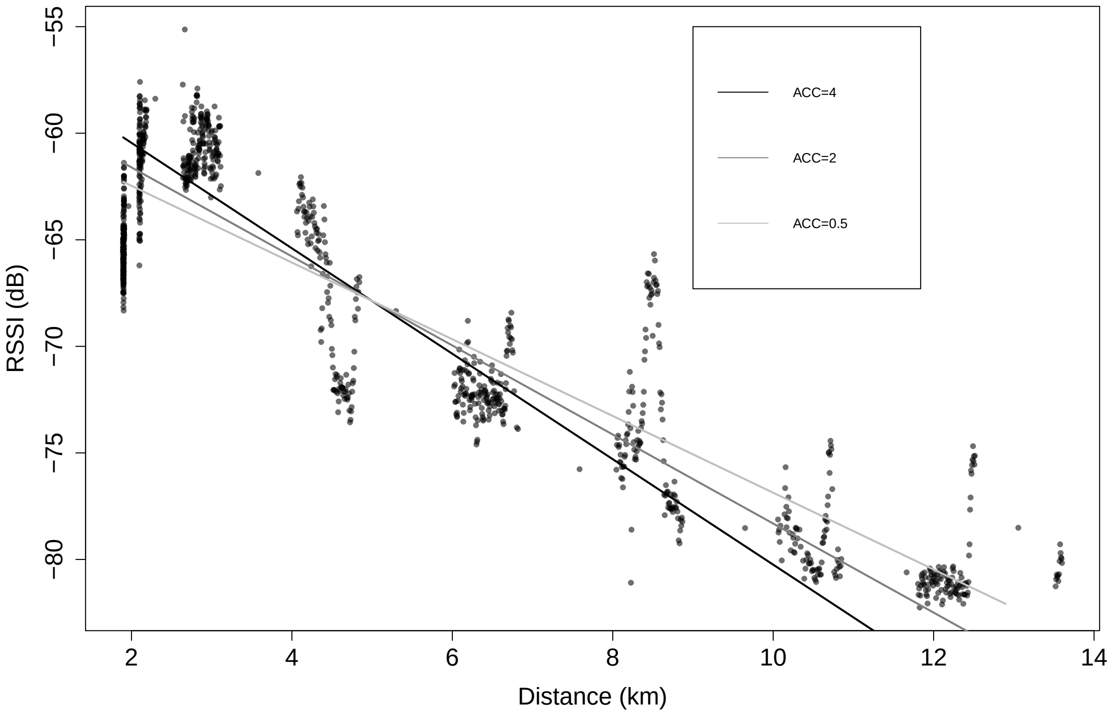


Figure 6.22: RSSI as function of distance and predictions of different vessel antenna's acceleration values.

Analyzing Figure 6.22, the predictions indicates that the sea oscillation has an impact on RSSI, that is the relationship between RSSI and distance differs with vessel's antenna acceleration. The vessel motion under greater oscillation results on more degradation of the communication signal.

6.3 Best predictors

As stated in Section 3.4, when there are several predictors, we need a strategy to select the best predictors to use in a regression analysis. Hence, a study was made in order to perceive which predictors have the most influence in this context. With three predictors, there are $2^3 = 8$ possible models. All 8 possible models were fitted and the results are summarized in Table 6.3. A "1" in the predictors column means that the predictor was included in the model, and a "0" indicates that the predictor was not included in the model. The results have been arranged according to AICc and consequently the best models are disposed at the top of the table.

The best model includes all the three predictors. Nevertheless, a closer look exposes some interesting features. There is a clear gap between the models in the first four rows and the ones below. This indicates that the Distance is the most important variable. Also, the first two rows

Table 6.3: Models for the RSSI variable with 3 predictors.

Predictors			Measures					Rank average
Distance	ACC	Noise	CV	AIC	AICc	BIC	AdjR2	
1	1	1	9.151	2449.259	2449.314	2474.297	0.821	2.8
1	1	0	9.154	2449.749	2449.786	2469.780	0.820	2.8
1	0	0	9.231	2458.258	2458.280	2473.281	0.819	3.4
1	0	1	9.246	2460.062	2460.098	2480.092	0.818	4.2
0	1	1	32.047	3834.343	3834.379	3854.373	0.372	4.8
0	0	1	32.287	3842.444	3842.466	3857.467	0.367	5.4
0	1	0	45.202	4213.486	4213.507	4228.508	0.115	6
0	0	0	51.022	4347.142	4347.153	4357.157	0.000	6.6

have almost identical values of CV, AIC and AICc, which means that we could possibly drop the Noise and achieve very similar results.

Further more, a rank average was made in order to find out which would be the best model to apply. It turns out that the rank emphasize what was said in the previous paragraph, that possibly the Noise values may be dropped, since the first two models have exactly the same rank average.

The best model selected by both rank average and measurements is given by

$$RSSI = \beta_0 + \beta_1 Distance + \beta_2 ACC + \beta_3 NOISE. \tag{6.4}$$

Table 6.4 demonstrates that applying the model with the three predictors, some variables maintain their influence on the RSSI, while others become less influential. That is, comparing this model with the models shown in Section 6.2.2, it is possible to understand that the distance has the same influence on the RSSI, while the vessel's antenna acceleration and the noise have a much smaller influence. Once again it is reinforced that the variable that has the greatest influence on the communication channel is the distance.

Table 6.4: Coefficients values for the best model.

Coefficient	Value
β_0	-51.21 dB
β_1 (Distance)	-1.80 dB
β_2 (ACC)	-0.56 dB
β_3 (NOISE)	0.08 dB

6.4 Propagation models

As stated in Section 4.5, there are two propagation models that can fit with the tests performed, taking into account the maritime environment. In this section, since the main goal was to understand which model of propagation is best suited, we used the data related to both directions, normal and rev, for better perception.

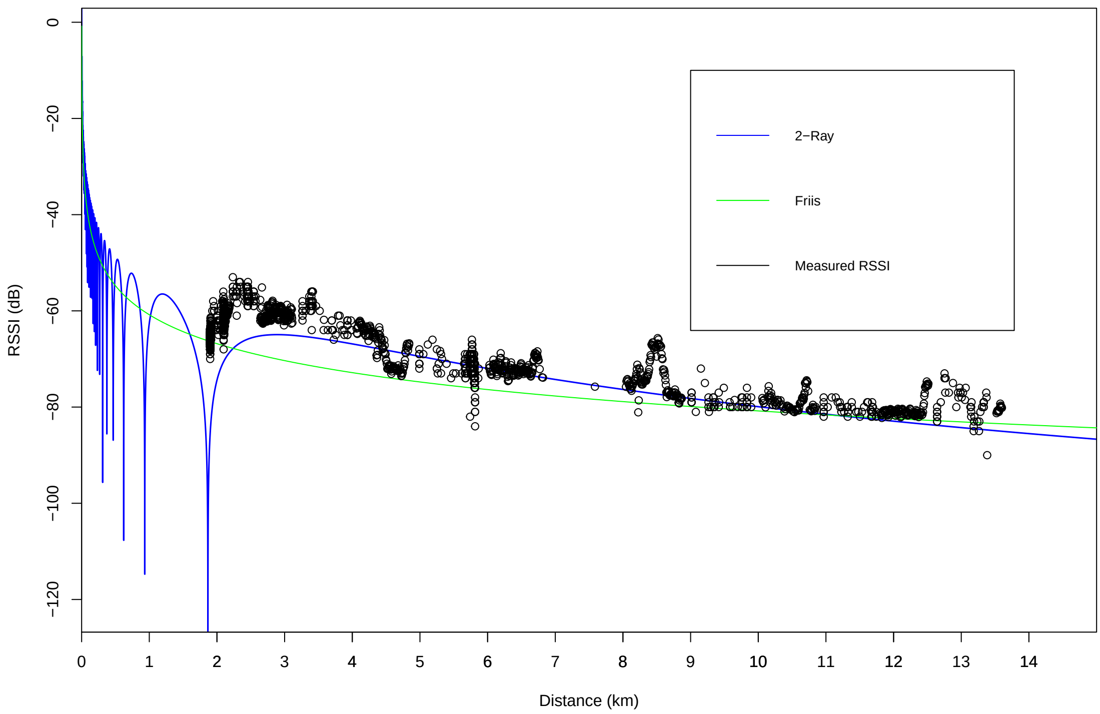


Figure 6.23: Propagation models: simulation vs measured signal.

The signal is no longer detected from about 13.5 km, due to difficulties encountered as a result of the testbed's lack of processing capacity. As the vessel moves away from the coast, the received signal strength has some values that deviate from the theoretical ones, this could be due to several reasons, such as temporary loss of signal, interference in the communication, the constant variation of the height of the receiving antenna in relation to the sea level, etc. But, even taking all this reasons in mind, considering Figure 6.23 it is possible to conclude that the measured received signal fits better in the 2-ray model.

Chapter 7

Conclusion and Future Work

7.1 Achievement of objectives

The focus of this dissertation was to search for and identify relationships between data communication variables and external variables that may occur in large amounts of information related to the transmission of data in maritime environments. Alongside this, other important objective was to develop an automated and intuitive framework that can be used in similar future projects that might arise.

Due to the difficulties that occurred during the tests performed, the amount of data for possible analysis was smaller than expected, thus restricting the analysis made.

Initially, it was necessary to perceive the network topology, the hardware and the software used in the project testbed and a theoretical study was made concerning which propagation models are better suited for this context.

Afterwards, a data processing was made, in order to study the variables presented in the logs, organizing the data, sample the data, perceive the route made by the vessel on different days and attempt to identifying relationships between communication variables and external variables.

After the data processing is completed, a study were made focused on identifying possible relationships between data communication variables and external variables, such as environmental variables, direction and location in two different scenarios. Later, a data analytics technique was applied and taking into account the data generated, it was demonstrated which propagation model fits better. In Chapter 2, several techniques to analyze the data were described, but taking in account the data provided, the most appropriate and successful technique was regression.

In a conclusive way, and having in mind the amount of data resulting from the tests, it was possible to make an analysis of the Wi-Fi communication in maritime environment, in order to understand some aspects that could affect the power of the received signal and factors to which the vessel is subject in this kind of environment. A framework has also been developed in a way that is easily used for further analysis on similar projects, including MareCom tests that will take place this year with improvements in terms of hardware and software.

7.2 Conclusion

The performed tests have proved the viability of using an IEEE 802.11 network supported by low-cost off-the-shelf hardware as a means of providing Internet access in a maritime environment several kilometers from the shore. However, the throughput achieved at this stage is clearly insufficient to accommodate the traffic generated by the applications, such as, a video call that would consume the entire available bandwidth, making it impossible to use any other applications simultaneously. Also, with the increase of the number of vessels served by a single AP, the capacity of the channel would degrade in such a way that in the limit any type of communication would become impracticable.

Taking into account the generated data and the used antennas, with the data analytics made in this dissertation, it was concluded that the maritime environment has a significant impact on the stability of the communications signals, being one of the main reasons for the variability of the signal the oscillation of the vessels. Applying linear regression, it was possible to understand several aspects in this context, such as, the relation between some external variables and communication variables, the influence of these variables on the quality of the transmission, which are the external variables that influences more the received power by the vessel and perform a few predictions of the received power taking into consideration certain variables.

7.3 Future work

Several important upgrades are planned to be made in the MareCom testbed conditions, as for example, the increase of the hardware processing/storage capacity, the increase of the throughput rate and the use of other type of antennas (multi-sector). The work developed in this dissertation may take advantage of these improvements.

In the future, more campaigns will be carried out to collect more data in order to validate the results with a large set of data sets and more diversity in terms of external conditions.

With the generation of more and varied data, in terms of external conditions, frequency of operation, etc., it may be possible to apply other alternative data analytics techniques, such as random forests, in order to better understand the Wi-Fi communication in the maritime environment.

Appendix A

Scripts

This appendix intends to present some of the scripts developed in R and python and described in Chapter 5.

A.1 HORST logs gathering script

Listing A.1: HORST logs gathering script

```
setwd("C:/Users/Jose Carvalho/Desktop/Dados_tratados/Horst/17_11/Normal/")
path <-"C:/Users/Jose Carvalho/Desktop/Dados_tratados/Horst/17_11/Normal/"
 files <- list . files (path=path, pattern ="*.csv")

`17112016_SRC1_total` <- data.frame(NULL)
`17112016_SRC2_total` <- data.frame(NULL)

for( file in files )
{
  perpos <- which( strsplit ( file , "") [[1]]==".")
   datafiles <- read.csv( file ,sep=",",head=FALSE,row.names=NULL)

   datafiles = datafiles [-1,]   #remove a linha que tem o header

  colnames( datafiles ) <- c("TIME","TYPE","SRC","DST","BSS","PKT_TYPE","RSSI","RATE IDX","RETRY","PKT_LENGTH","RATE","NOISE","ANTENNA IDX","RSSI1","ANTENNA IDX","RSSI2","ANTENNA IDX","RSSI3","ESSID","MODE","FREQ","CHANNEL","SEQ NO","IP SRC","IP DST","WLAN TSF","LOC TIME","ACCX","ACCY","ACCZ","GIRX","GIRY","GIRZ","MAGX","MAGY","MAGZ","LAT","LONG","ALT")

  #Valores relativos (ACC, GIR, MAG)
```

```
  datafiles $ACCX <− as.numeric(as.character( datafiles $ACCX))
  datafiles $ACCX<− ave(datafiles$ACCX, FUN=function(x) c(0, diff(x)))

  datafiles $ACCY <− as.numeric(as.character( datafiles $ACCY))
  datafiles $ACCY<− ave(datafiles$ACCY, FUN=function(x) c(0, diff(x)))

  datafiles $ACCZ <− as.numeric(as.character( datafiles $ACCZ))
  datafiles $ACCZ<− ave(datafiles$ACCZ, FUN=function(x) c(0, diff(x)))

  datafiles $GIRX <− as.numeric(as.character( datafiles $GIRX))
  datafiles $GIRX<− ave(datafiles$GIRX, FUN=function(x) c(0, diff (x)))

  datafiles $GIRY <− as.numeric(as. character ( datafiles $GIRY))
  datafiles $GIRY<− ave(datafiles$GIRY, FUN=function(x) c(0, diff (x)))

  datafiles $GIRZ <− as.numeric(as. character ( datafiles $GIRZ))
  datafiles $GIRZ<− ave(datafiles$GIRZ, FUN=function(x) c(0, diff (x)))

  datafiles $MAGX <− as.numeric(as.character( datafiles $MAGX))
  datafiles $MAGX<− ave(datafiles$MAGX, FUN=function(x) c(0, diff(x)))

  datafiles $MAGY <− as.numeric(as.character( datafiles $MAGY))
  datafiles $MAGY<− ave(datafiles$MAGY, FUN=function(x) c(0, diff(x)))

  datafiles $MAGZ <− as.numeric(as.character( datafiles $MAGZ))
  datafiles $MAGZ <− ave(datafiles$MAGZ, FUN=function(x) c(0, diff(x)))

  datafiles <− datafiles [−nrow( datafiles ) ,] #remove a linha que fica com NA

  datafiles _qdata <− subset( datafiles , TYPE==" QDATA")

  datafiles _qdata_SRC1 <− subset( datafiles _qdata, SRC==" 00:30:1a:46:1e:c1")

  datafiles _qdata_SRC2 <− subset( datafiles _qdata, SRC==" 00:30:1a:46:1e:c2")

# juntar horsts relativos a SRC1
‘17112016_SRC1_total‘ <− rbind(‘17112016_SRC1_total‘,‘ datafiles _qdata_SRC1‘)
# juntar horsts relativos a SRC2
‘17112016_SRC2_total‘ <− rbind(‘17112016_SRC2_total‘,‘ datafiles _qdata_SRC2‘)

```

```
  #Apagar ficheiros ja usados
  rm( datafiles )
  rm(‘ datafiles _qdata ‘)
  rm(‘ datafiles _qdata_SRC1‘)
  rm(‘ datafiles _qdata_SRC2‘)
}
```

A.2 Iperf logs gathering script

Listing A.2: Iperf logs gathering script

```
#Cenario ESTATICO

setwd("C:/Users/Jose Carvalho/Desktop/Dados_tratados/ Iperf /17_11/Normal/Estatico/")
path <−"C:/Users/Jose Carvalho/Desktop/Dados_tratados/ Iperf /17_11/Normal/Estatico/"
 files <− list . files (path=path, pattern ="∗.csv")

‘17112016_iperf_ estatico ‘ <− data.frame(NULL)

for( file in files )
{
  perpos <− which( strsplit ( file , "") [[1]]==".")
   datafiles <− read.csv( file ,sep="",head=FALSE,row.names=NULL)

   datafiles = datafiles [−248:−254,]

   datafiles = datafiles [−1:−6,]

   datafiles $V1 <− NULL
   datafiles $V2 <− NULL
   datafiles $V4 <− NULL
   datafiles $V6 <− NULL
   datafiles $V8 <− NULL
   datafiles $V10 <− NULL

  colnames( datafiles ) <− c(" Interval (sec)" ," Transfer (KBytes)","Bandwidth(Mbits/sec)" ,"
        Jitter (ms)","Lost/ Total _Datagrams")

   datafiles _= datafiles [seq (2, nrow( datafiles ), 2), ]

```

```
  toDel <- seq(1, nrow( datafiles ), 2)

  tmp <- datafiles [ toDel ,]

  # juntar iperfs
  `17112016_iperf_estatico` <- rbind(`17112016_iperf_estatico`,`tmp`)

  #Apagar ficheiros ja usados
  rm( datafiles )
  rm( datafiles_)
  rm(tmp)

}

#remove as ultimas linhas que nao sao necessarias, visto que o ultimo ficheiro nao tem 120 segundos
`17112016_iperf_estatico` <- `17112016_iperf_estatico`[-nrow(`17112016_iperf_estatico`),]

save('17112016_iperf_estatico', file="17112016_iperf_estatico.RData")

#Cenario MOVIMENTO
setwd("C:/Users/Jose Carvalho/Desktop/Dados_tratados/Iperf/17_11/Normal/Movimento/")
path <-"C:/Users/Jose Carvalho/Desktop/Dados_tratados/Iperf/17_11/Normal/Movimento/"
 files <- list.files(path=path, pattern="*.csv")

`17112016_iperf_movimento` <- data.frame(NULL)

for( file in files )
{
  perpos <- which( strsplit( file , "")[[1]]==".")
   datafiles <- read.csv( file ,sep="",head=FALSE,row.names=NULL)

   datafiles = datafiles[-248:-254,]

   datafiles = datafiles[-1:-6,]

   datafiles$V1 <- NULL
   datafiles$V2 <- NULL
   datafiles$V4 <- NULL
```

```
  datafiles $V6 <- NULL
  datafiles $V8 <- NULL
  datafiles $V10 <- NULL


 colnames( datafiles ) <- c(" Interval (sec)"," Transfer (KBytes)","Bandwidth(Mbits/sec)","
      Jitter (ms)","Lost/ Total_Datagrams")

  datafiles _= datafiles [seq(2, nrow( datafiles ), 2), ]

 toDel <- seq(1, nrow( datafiles ), 2)

 tmp <- datafiles [ toDel ,]

 # juntar iperfs
 `17112016_iperf_movimento` <- rbind(`17112016_iperf_movimento`,`tmp`)

 #Apagar ficheiros ja usados
 rm( datafiles )
 rm( datafiles _)
 rm(tmp)

}

#remove as ultimas linhas que nao sao necessarias , visto que o ultimo ficheiro nao tem
     120 segundos
`17112016_iperf_movimento` <- `17112016_iperf_movimento`[-nrow(`17112016_iperf_
    movimento`),]

save('17112016_iperf_movimento',file="17112016_iperf_movimento.RData")
```

A.3 Coordinates converter script

Listing A.3: Coordinates converter script

```
#Load files
setwd("C:/Users/Jose Carvalho/Desktop/Dados_tratados/Horst/17_11/Normal/Sampled Data"
    )
path <-"C:/Users/Jose Carvalho/Desktop/Dados_tratados/Horst/17_11/Normal/Sampled Data
    "
```

```
load( file ="17112016_SRC1_movimento_mean152.RData")

#COnverter as coordenadas de Degrees, Decimal Minutes para Decimal Degrees

myvars <− c("RSSI","LAT","LONG","ALT")

tmp <− ‘17112016_SRC1_movimento_mean152‘[myvars]

coord <− tmp

for (row in 1:nrow(tmp))
{
  tmp$DEG_LAT = 41
  tmp$MIN_LAT = −(4100−tmp$LAT)
  coord$LAT = tmp$DEG_LAT+(tmp$MIN_LAT/60)

  tmp$DEG_LONG = 8
  tmp$MIN_LONG = −(800−tmp$LONG)
  coord$LONG = −(tmp$DEG_LONG+(tmp$MIN_LONG/60))
}

write . table (coord, file ="coords_17112016_SRC1_movimento_mean152.csv",row.names=
    FALSE, na="",col.names=TRUE, sep=",")

#A partir daqui a tabela esta apta para introduzir no script (kml.py)

#Calculo da distancia

for (row in 1:nrow(coord))
{
  coord$DISTANCE = acos( sin(41.201312∗pi/180)∗sin(coord$LAT∗pi/180) + cos(41.201312∗
      pi/180)∗cos(coord$LAT∗pi/180)∗cos(coord$LONG∗pi/180−(−8.712188)∗pi/180) ) ∗
      6371000
  coord$"DISTANCE(km)"= coord$DISTANCE∗10^(−3)
}
```

A.4 KML file generator script

Listing A.4: KML file generator script

```
import csv
#Colocar o nome do ficheiro
fname = raw_input("Enter file name WITHOUT extension: ")
data = csv.reader(open(fname + '.csv'), delimiter = ',')
#Avancar a 1 linha (header)
data.next()
#Abrir o ficheiro para leitura
f = open('coords.kml', 'w')

#Gerar ficheiro KML
f.write("<?xml version='1.0' encoding='UTF-8'?>\n")
f.write("<kml xmlns='http://earth.google.com/kml/2.1'>\n")
f.write("<Document>\n")
f.write("   <name>" + fname + '.kml' +"</name>\n")
for row in data:
    f.write("   <Placemark>\n")
    f.write("       <name>" + str(row[0]) + "</name>\n")
    f.write("       <Point>\n")
    f.write("           <coordinates>" + str(row[2]) + "," + str(row[1]) + "," + str(row[3]) + "</coordinates>\n")
    f.write("       </Point>\n")
    f.write("   </Placemark>\n")
f.write("</Document>\n")
f.write("</kml>\n")
print "File Created. "
print "Press ENTER to exit. "
raw_input()
f.close()
```

Bibliography

[1] David Kidston and Thomas Kunz. Challenges and opportunities in managing maritime networks. *IEEE Communications Magazine*, 46(10), 2008.

[2] INESCTEC. Projeto marecom: alternativa ao satélite e aos rádios vhf em alto mar. URL: https://www.inesctec.pt/ctm/noticias-eventos/noticias/projeto-marecom-alternativa-ao-satelite-e-aos-radios-vhf-em-alto-mar/. [Online; accessed 28-June-2017].

[3] Chee-Wei Ang and Su Wen. Signal strength sensitivity and its effects on routing in maritime wireless networks. In *33rd IEEE Conference on Local Computer Networks, LCN 2008*, pages 192–199. IEEE, 2008.

[4] Sourangsu Banerji and Rahul Singha Chowdhury. On IEEE 802.11: Wireless LAN Technology. *International Journal of Mobile Network Communications and Telematics*, 2013.

[5] Nayarasi. Cwap – 802.11 ctrl : RTS/CTS. URL: https://mrncciew.com/2014/10/26/cwap-802-11-ctrl-rtscts. [Online; accessed 13-June-2017].

[6] Jochen H Schiller. *Mobile communications*. Pearson Education, 2003.

[7] Dionysios Skordoulis, Qiang Ni, Hsiao-Hwa Chen, Adrian P Stephens, Changwen Liu, and Abbas Jamalipour. IEEE 802.11n MAC frame aggregation mechanisms for next-generation high-throughput WLANs. *IEEE Wireless Communications*, 15(1), 2008.

[8] Samir Raiyani. LTE. URL: https://www2.dolcera.com/wiki/index.php. [Online; accessed 12-June-2017].

[9] Aerohive Network. 802.11n Primer. URL: https://www.infosec-cloud.com/wp-content/uploads/2013/01/Aerohive-Whitepaper-802-11n_Technology_Primer.pdf. [Online; accessed 13-June-2017].

[10] Josip Lorincz and Dinko Begusic. Physical layer analysis of emerging IEEE 802.11n WLAN standard. In *8th International Conference Advanced Communication Technology, 2006. ICACT 2006*, volume 1, pages 189–194. IEEE, 2006.

[11] Ian F Akyildiz and Xudong Wang. *Wireless mesh networks*, volume 3. John Wiley & Sons, 2009.

[12] M Series. Maritime broadband wireless mesh networks. 2011.

[13] Jeffrey G Andrews, Arunabha Ghosh, and Rias Muhamed. *Fundamentals of WiMAX: understanding broadband wireless networking*. Pearson Education, 2007.

[14] J Joe, SK Hazra, SH Toh, WM Tan, J Shankar, Vinh Dien Hoang, and Masayuki Fujise. Path loss measurements in sea port for WiMAX. In *Wireless Communications and Networking Conference, 2007. WCNC 2007*, pages 1871–1876. IEEE, 2007.

[15] Jaya Shankar Pathmasuntharam, Joe Jurianto, Peng-Yong Kong, Yu Ge, Mingtou Zhou, and Ryu Miura. High speed maritime ship-to-ship/shore mesh networks. In *7th International Conference on ITS Telecommunications, 2007. ITST'07*, pages 1–6. IEEE, 2007.

[16] Ming-Tuo Zhou, Vinh Dien Hoang, Hiroshi Harada, Jaya Shankar Pathmasuntharam, Haiguang Wang, Peng-Yong Kong, Chee-Wei Ang, Yu Ge, and Su Wen. Triton: High-speed maritime wireless mesh network. *IEEE Wireless Communications*, 20:134–142, 2013.

[17] Rosario Giuseppe Garroppo, Stefano Giordano, and Davide Iacono. Experimental and simulation study of a WiMAX system in the sea port scenario. In *IEEE International Conference on Communications, 2009. ICC'09*, pages 1–5. IEEE, 2009.

[18] Vijay Garg. *Wireless communications & networking*. Morgan Kaufmann, 2010.

[19] Yanwen Wang, Ivan G Guardiola, and Xiaoling Wu. RSSI and LQI data clustering techniques to determine the number of nodes in wireless sensor networks. *International Journal of Distributed Sensor Networks*, 2014.

[20] Saravanan Kandasamy, Ricardo Morla, Patrícia Ramos, and Manuel Ricardo. Predicting throughput in ieee 802.11 based wireless networks using directional antenna. *Wireless Networks*, pages 1–18, 2017.

[21] Rob J Hyndman, Earo Wang, and Nikolay Laptev. Large-scale unusual time series detection. In *IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 14–17. IEEE, 2015.

[22] John O Rawlings, Sastry G Pantula, and David A Dickey. *Applied regression analysis: a research tool*. Springer Science & Business Media, 2001.

[23] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2014.

[24] M Subba Rao and B Eswara Reddy. Comparative analysis of pattern recognition methods: An overview. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(3):385–390, 2011.

[25] Douglas M Hawkins. *Identification of outliers*. Springer, 1980.

[26] Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

[27] Joyce Jackson. Data mining; a conceptual overview. *Communications of the Association for Information Systems*, 8(1):19, 2002.

[28] George Athanasopoulos Rob J Hyndman. Forecasting: Principles and Practice. URL: http://otexts.org/fpp2/. [Online; accessed 7-December-2017].

[29] Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1:14–23, 2011.

[30] OpenWrt. URL: https://openwrt.org/. [Online; accessed 28-December-2017].

[31] Windguru. . URL: https://www.windguru.cz/. [Online; accessed 3-December-2017].

[32] Mário Jorge de Matos Lopes. Comunicações marítimas Wi-Fi usando a banda 4.8 GHz. 2013.

[33] HORST. URL: https://github.com/br101/horst. [Online; accessed 17-December-2017].

[34] Iperf. URL: https://iperf.fr/. [Online; accessed 17-December-2017].